



ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES
SCHOOL OF INFORMATION SCIENCE

**Data Quality Improvement Model for Awash Insurance's
Information System**

By: Naol Gemedja Jaleta

ID: GSE/4916/14

September, 2024

ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES
SCHOOL OF INFORMATION SCIENCE

**Data Quality Improvement Model for Awash Insurance's
Information System**

A Thesis Submitted to school of Information Science
of Addis Ababa University in Partial Fulfillment of the
Requirements for the Degree of Master of Science in Information
Science and Systems (Information Science)

By: Naol Gemed Jaleta

Advisor: Dr.Melkamu Beyene

ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES
SCHOOL OF INFORMATION SCIENCE

**Data Quality Improvement Model for Awash Insurance's
Information System**

By: Naol Gemed Jaleta

Approval

<u>Name</u>	<u>Title</u>	<u>Signature</u>	<u>Date</u>
Dr. Melkamu Beyene	Advisor	_____	_____
Dr. Dereje Teferi	Examiner	_____	_____
Dr. Tibebe Beshah	Examiner	_____	_____

DECLARATION

I declare that the thesis “*DATA QUALITY IMPROVEMENT MODEL FOR AWASH INSURANCE’S INFORMATION SYSTEM*” is my original work and has not been presented for a degree in any other university.

Name of the Student

Naol Gemed

Signature

Date

This thesis has been submitted for examination with my approval as university advisor.

Dr. Melkamu Beyene

ACKNOWLEDGMENTS

I would like to first and foremost thank Almighty God for his grace, wisdom, and guidance throughout this journey. Without his strength, this work would not have been possible.

My deepest gratitude goes to my advisor, Dr. Melkamu Beyene, for his invaluable guidance, encouragement, and expertise. Your support has been instrumental in shaping this research.

I am profoundly grateful to my family for their unwavering love and encouragement, and also, I would like to extend my thanks to the involved Awash Insurance staff for their assistance and support throughout my research.

Finally, to my wonderful wife, Hanna Kiflu, I am forever grateful for your love, patience, and unwavering support. Your strength and encouragement have been my foundation, and this achievement is as much yours as it is mine.

Thank you all for your contributions to this work.

Contents

ACKNOWLEDGMENTS	iv
Chapter 1: Introduction	1
1.1. Background	1
1.2. Study Motivation.....	2
1.3. Statement of the Problem	3
1.3.1. Research Questions.....	5
1.4. Objective of the thesis	5
1.5. Specific Objectives of the study	6
1.6. Scope of the Research	6
1.7. Significance of the study	6
Chapter 2: Literature Review	8
2.1. Overview of data Quality in Information Systems.....	8
2.2. Data Quality Dimensions	9
2.2.1. Intrinsic Dimensions.....	10
2.2.2. Contextual Dimensions.....	11
2.3. Data Quality in Insurance IS	12
2.4. Data Quality Management Techniques	12
2.4.1. Total DQ Management (TDQM).....	13
2.4.2. Information Integrity Methodology (IIM).....	14
2.4.3. AIM Quality model (AIMQ)	14
2.4.5. Complete DQ Management (CDQM)	15
2.4.6. Six Sigma.....	16
2.4.7. Master Data Management (MDM)	17
2.4.8. Data Stewardship Model	17
2.5. Review of Related Works	18
2.5.1. Data Quality improvement in Insurance Industry.....	20
Chapter 3: Research Methodology.....	22
3.1. Description of study's research process methodology	23
3.1.1. Problem identification and Definition	23
3.1.2. Definition of the objectives for a solution	24
3.1.3. Design and Development.....	25
3.1.4. Demonstration	25

3.1.5. Evaluation	26
3.1.6. Communication	26
3.2. Data Collection Methods.....	26
3.2.1. Data Collection	26
3.2.2. Sampling Strategy.....	28
3.2.3. Data Analysis Method and Finding	28
3.2.4. Ethical Considerations	30
Chapter 4: Design and Development.	31
4.1. Requirements Specification.....	31
4.1.1. Approach of requirements Specification.	31
4.1.2. Functional Requirement.	32
4.1.3. Non Functional Requirement.....	32
4.1.4. Stakeholder Identification.....	33
4.2. Description of Literature Review process.....	34
4.2.1. Approach of Literature Selection for Solution Design	35
4.2.2. Summary of Interview result	35
4.2.3. Summary of Document review analysis.....	36
4.3. Analysis of Existing DQ Management methodology and Model customization.....	36
4.3.1. Detail description and Analysis of DQ Management methodology.....	36
4.3.2. Conclusion of the Analysis of the Methodologies and model.....	38
4.4. Design of the customized Model.....	38
4.4.1. Proposed Insurance Data Quality Improvement Process Model	41
4.4.2. Description of the Proposed Data Quality Improvement Process Model.....	42
Chapter 5: Demonstration and Evaluation.....	48
5.1. Overview	48
5.2. Proposed Model description.....	48
5.3. Implementation Plan	49
5.4. Execution of Demonstration for Data Quality Improvement.....	52
5.5. Conclusion of the Demonstration process.....	57
5.6. Evaluation process.....	58
Chapter 6: Conclusion and Recommendation.....	61
6.1. Conclusion.....	61
6.2. Recommendation.....	61
References	63

Appendix.....	68
Appendix I: Interview question Checklist.....	68
Appendix II: Checklist for Organizational document analysis.	69
Appendix III: Search terms used for literature review	70
Appendix IV: Awash Insurance Organizational structure	71
<i>(Used to understand the company's business process)</i>	71
Appendix V: One of System Manual used for Document Analysis	72

List of Tables

Table 1 : Data Quality Management methodologies analysis	36
Table 2:Customer Information data with Data Quality Issue	53

List of figures

Figure 1: Dimensions of Data Quality	10
Figure 2: A Summary of the TDQM methodology created by Battini et al. (2009).....	14
Figure 3: Six Sigma's DMADV Process	17
Figure 4: Design Science Research's Process model (Peffer et al., 2007).....	23
Figure 5: Proposed Insurance Data Quality Improvement Process Model.....	42
Figure 6: python script code for Instance duplication removal	55

List of Acronyms

AIC	Awash Insurance Company
AIMQ	Information Quality Assessment and Improvement Method
CDQM	Complete DQ Management
CIF	Customer Information File
DK	design knowledge
DMADV	(Define, Measure, Analyze, Design and Verify)
DMAIC	(Define, Measure, Analyze Improve and Control)
DQ	Data Quality
DSR	Design Science Research
DSRM	Design Science Research Methodology
ERP	Enterprise Resource Planning.
EU	European Union
FAIR	(find ability, accessibility, interoperability, reusability)
FR	Functional requirements
GIIS	General Insurance Information System.
IDQIP	Insurance Data Quality Improvement Process
IP	Information Product
IQ	Information Quality
IS	Information System.
ISO	International Organization for Standardization
LHIIS	Life and Health Insurance Information System
MDM	Master Data Management
ML	Machine learning
NFRs	Non-functional requirements
NLP	Natural Language Processing
PSP	Product and Service Performance
TDQM	Total Data Quality Management Methodology
TQM	Total Quality Management

Abstract

The study analyzes and improves data quality issues in the information system of Awash Insurance S.C. through providing a comprehensive model known as Insurance Data Quality Improvement Process Model (IDQIPM). The approach contributes to operational efficiency and decision-making reliability of the company (AIC) by addressing data quality problems such as instance duplication, data value inconsistency, inaccuracy, and incompleteness issues of its Information system data. Design Science Research Methodology with qualitative approach is employed in this study. A data quality assurance model that can enhance data quality within Insurance's underwriting operation, one of the core business processes of the company and that faces most of data quality problems is proposed. The model's effectiveness has been attempted to be shown in demonstration section of this paper which brought about encouraging results in terms of enhancing data quality within Awash Insurance S.C.'s insurance information system. However, future research might improve the model so as to enhance various data quality dimensions which is not addressed in this study.

Chapter 1: Introduction

1.1. Background

Nowadays, activities and decisions making in an organization is based on data and information obtained from data analysis, which provides various services for constructing reliable and accurate process[1]. Information System is a technology that has been applied in every aspect of Organizations. Most of the organization's business processes are increasingly growing to be reliant on this technology. Financial Organizations like Insurance companies are one of those organizations that use the Information system technology for their business process.

Most of Insurance companies have been implemented their own Information system to support for their respective overall organizational business activities.

Awash Insurance Company is one of private Insurance Company in Ethiopia which was launched following the liberalization of the financial sector in Ethiopia. The company (AIC) manages its business operations using an integrated information system. The system comprises three major subsystems: General Insurance Information System (GIIS) which is used to process and manage non-life insurance operations, Life and Health Insurance Information System (LHIIS) to process and manage life and health insurance businesses and Enterprise Resource Planning (ERP) for managing company's resources.

The insurance information system uses data that is acquired from different sources and stored in the system database to provide the necessary Information for Organizational business process; For instance Awash Insurance Company's customers are required to provide their personal and industry-specific information, such as their names, residences, phone numbers, medical histories, accident records, and legal related data. The integrated information system stores this information as alphabetic or numerical data in its database.

In the case of Awash Insurance companies' integrated insurance information systems, most of the data are user entry. So based on the entered data into the system, every Organizational business process will be processed.

In Insurance core business processes, to determine the eligibility for customers for Policy (Insurance cover), to decide the premium (Insurance service cost) amount to be paid, to decide claims type and amount to be settled, to analyze the data for decision making and for customer

satisfaction, the company highly depends on the data. So when we think of data, data quality issue should be one of the main concerns of organization because the concern has high significant effect on the overall business process of the company

Based on a report by Laatikainen & Niemi (2012) in Finland it is estimated that costs due to data quality issues are around 10 billion Euros per year. This has caused companies to not trust their data and causes difficulties to make business decisions based on poor data [2].

The concept of data quality is widely interpretable. A common definition given to data quality is “*Fitness for use*”, which implies the meaning varies depending on the use of the data (Vincent.B, 2023).

Corporations are losing significant amount of money, and missing important opportunities all the time because they operate on information derived from inaccurate data. (Olson, J. E. (2003)).

Generally this thesis research aims to analyze and improves data quality issues of insurance information system (particularly for Awash Insurance), that has significant effect on decision making process of the company through providing an appropriate process model.

1.2. Study Motivation

The growing importance of data quality in the insurance sector, particularly within the Awash Insurance Information System, underscores the need for effective data management practices.

Bad data quality costs organizations an average of \$12.9 million per year (Gartner, 2021). With increasing competition and regulatory requirements, the ability to rely on accurate and timely data is crucial for decision-making and operational efficiency. This study aims to address the challenges of data quality by developing a comprehensive improvement model tailored to the specific needs of Awash Insurance.

The motivation for this research stems from observed data redundancy, inconsistencies and inaccuracies in existing data, which not only hinder the performance of the information system but also impact customer satisfaction and trust. By focusing on data quality enhancement, this study seeks to contribute to the overall effectiveness of the insurance services offered by Awash, fostering better service delivery and risk management.

Furthermore, the findings from this research may serve as a framework for other organizations in the insurance sector facing similar data quality issues, promoting a culture of data-driven decision-making and continuous improvement.

1.3. Statement of the Problem

In today's highly competitive insurance industry, the accuracy and reliability of data are paramount to operational success and customer satisfaction. Awash Insurance, like many other organizations, relies heavily on its information system for managing a range of critical processes, including policy issuance, claims processing, risk assessment, and customer service. However, recent assessments and feedback have highlighted several issues related to data quality within their system.

Data accuracy and completeness are fundamental to effective decision-making and operational efficiency. At Awash Insurance, there have been numerous instances of inaccurate or incomplete data entries, which have led to errors in claims processing, incorrect risk evaluations, and delays in policy management. These inaccuracies not only affect the company's operational efficiency but also diminish the trust and satisfaction of customers.

The presence of duplicate records and redundant information within the Awash Insurance information system has been identified as a significant problem. Such issues lead to inefficiencies in data retrieval, increased storage costs, and potential confusion among staff when accessing client information or processing claims.

The absence of data consistency across departments and systems also hinders decision-making reliability. Inadequate data validation and improvement processes also contribute to erroneous data entry and complications in data processing and reporting. Addressing these issues is crucial for improving overall performance and maintaining a competitive edge.

Most of the data coming as input to the systems of the insurance companies are either from the Customer care(service) or from the customer himself, from information on the company's websites, portals or by mailing in the data(Panwar.M, 2015).

As human beings, we often make errors that can lead to misleading outcomes. In relation with data, mistakes are more possible to occur if not typed in properly [3].

For instance, the following hypothesis explains the data quality issues of the company's (AIC) Information system.

The same customer is registered in the system more than one time with a little bit spelling difference; this causes data duplication which has significant effect during claim processing operation and other operation's of the company's that uses customer related data.

The higher executive management of the company (AIC) once needed to know how many active customers it had for decision-making purposes. To do this, they conducted a detailed assessment with preferred stakeholders, including IT staff and data analysts. The results of the assessment revealed that only 45% of the total number of customers registered on the company's system software was active customers with policies. This indicates that the company had previously assumed that it had 55% more fake data than actual data.

In addition to the aforementioned hypothesis, data inconsistency and inaccuracy arise when users from various departments or branches of the business enter client contact information, such as email addresses and phone numbers, into the system in an erroneous or non-standard manner. On the other hand, AIC uses these contact details to engage with its customers regarding policy renewal and the claim procedure, both of which have a direct effect on the commercial performance of the organization.

After all, data quality problems in insurance company results inaccurate premium collection and claim settlement. In addition to this, it leads to regulatory penalties. Furthermore, it has significant impact on the company's decision-making process.

The company's system uses data type controlling mechanisms that helps to control data entry as per the system's data type configuration and Authorization workflow approach that allows managers to check and authorize transaction performed by operational users as a data quality improvement technique in its system. However, the approach doesn't address the overall data quality problems of the company's Information system.

Data quality improvement mechanisms can be categorized into several taxonomies based on various criteria. There are typically three types of Data quality improvement mechanisms Preventive, Detective and Corrective. Furthermore, a holistic approach combining these mechanisms across various dimensions is more essential for effective data quality improvement

for Insurance Organization's Information system like AI which requires a systematic process and continuous quality checks for ensuring better organizational quality.

Several scientific studies have been conducted on data quality issues in various firms, which aim to improve data quality of different data quality dimension using one of among the above mentioned improvement mechanisms.

This Thesis aims to follow a holistic approach, so as to address some of major identified data quality problems the company.

Data quality management methodologies are among holistic approaches to enhance data quality Management, which is the best solution to provide a comprehensive Data quality improvement of organizations like insurances because of the several reasons, however, it needs to be customized and improved so as to fit for specific organization's business process.

Methodologies are compared along several dimensions, including the methodological phases and steps, the strategies and techniques, the data quality dimensions, the types of data, and, finally, the types of information systems addressed by each methodology [53].

Following some recommendation of reviewed study works especially, that of DQ improvement study work in insurance industry and Awash Insurance Company's data quality problems and its improvement approach (like Data type Control and Authorization workflow) as gap, the purpose of this research is to improve data quality issues of the company through the use of an appropriate systematic data quality management strategy (DQ Improvement Process model).

1.3.1. Research Questions

1. What are critical data quality challenge that mostly impacts AIC's business operations and the existing data quality management best practices of this company?
2. What is an effective solution and best way for improving data quality in Awash Insurance Company's information system?
3. How will the proposed solution be evaluated?

1.4. Objective of the thesis

The goal of this thesis is to propose appropriate Data Quality Improvement Model for enhancing data quality in Awash Insurance Company's Information system.

1.5. Specific Objectives of the study

- a) Identify and analyze the critical data quality challenges impacting Awash Insurance Company's business operations and evaluate the current data quality management practices.
- b) Create a data quality enhancement model specifically for the information system of Awash Insurance Company.
- c) Design an evaluation framework for assessing the effectiveness of the proposed data quality improvement model.
- d) Provide recommendations for the implementation and continuous monitoring of the data quality improvement model.

1.6. Scope of the Research

This thesis aims to address critical data quality challenges that mostly impact the company's (AIC) business operations for instance data duplication, incompleteness, inconsistency, and inaccuracy within the Awash Insurance Information System. In addition to this, the research investigates how poor data quality affects decision-making processes in the organization (AI) and examines existing data quality management best practices of Awash Insurance.

The design and development of a data quality enhancement solution will also be subject to a scope definition. The paper's way of evaluation for a solution (model) as a designed artifact involves assessing its effectiveness, usability, and alignment with design goals. However, due to some constraints like finance, time and data availability, the researcher is forced to focus on more of the company's core business processes, such as Underwriting and Claims, in respective with above mentioned data quality dimensions and doesn't cover all units of the company's business information system.

1.7. Significance of the study

More than 20% of the Fortune 500 companies are insurance providers [3]. Insurance business is one of the fastest growing businesses in the world.

AIC is one of the growing insurance companies among financial institutions in Ethiopia that use integrated information systems that lead to data quality issues, which is one of the main concerns

of this organization. For this concern the study finding has the following significant effect not only on Awash Insurance but also for all Insurance companies in the country.

- It has a great contribution to the Ethiopian Insurance industry for the proper management data quality related issues.
- The finding of this research shall serve as a support for experts who have responsibility of Improving Information system data quality for financial institution especially for Insurance companies.
- Moreover, Higher education Academicians, IT Professional of the Company and Researcher on this area will get additional input and better experience through the body of knowledge.

Chapter 2: Literature Review

In this section some overview of concept related to Data, Insurance, Information System, Data quality, Data Quality Management and, some of applications of Machine learning and Natural Language Processing has been reviewed and briefly explained for the support of the study.

2.1. Overview of data Quality in Information Systems

We must first define data quality before we can investigate the DQ research that is currently accessible. Data Quality can be referred to be part of information quality [2]. Data Quality is a way of examining your data from different points of view (Blomqvist, N. (2019)).

Prior to delving into the explanation of data quality, we will first begin by characterizing data and quality independently. Data are values of qualitative and quantitative variables, belonging to a set of items [6]. It represents real world objects, in a format that can be stored, retrieved and elaborated by a software procedure [7]. Information is defined as data that processed to be useful [11].

Definition of *quality* varies and usually depends on the role of people who describes it [8]. "The standard of something as measured against other things of a similar kind; the degree of excellence of something" is what Oxford Dictionary defines as quality.

According to the Total Quality Management concept, quality is defined as "meeting the customer's needs" or "satisfying the customer" through a service-focused methodology.

In manufacturing, *"a measure of excellence or a state of being free from defects, deficiencies, and significant variations, brought about by the strict and consistent adherence to measurable and verifiable standards to achieve uniformity of output that satisfies specific customer or user requirements"* [9]. Lastly, quality is defined as "the totality of features and characteristics of a product or service that bears its ability to satisfy stated or implied needs" by ISO 8402-19868, which integrates the viewpoints of both the product and the service.

Just as with quality, it's challenging to come up with a common definition for DQ. Since data quality is a subjective concept, it might change based on the context in which it is used.

One of the common definitions of *data quality* is "fitness for use" which is also used to define quality in general; *"Data are of high quality if they are fit for their uses (by customers) in*

operations, decision-making, and planning. They are fit for use when they are free of defects and possess the needed features to complete the operation, make the decision, or complete the plan” [10]

2.2. Data Quality Dimensions

Understanding and recognizing the concept of data quality dimension is crucial for understanding and defining data quality. But what is a dimension of data quality?

Data Quality dimensions are a useful measurement approach to compare DQ levels across different systems over time [13].

There are many different sources describing and talking about the DQ dimensions; however, all of them almost followed the same approach of describing it, although it often refers to different levels and different data model elements [13].

David Loshin in [14] said that “Different dimensions are intended to represent different measurable aspects of DQ and are used in characterizing relevance across a set of application domains to monitor against the specified organizational standard of DQ”.

Data Quality dimensions are a helpful method of measurement for comparing DQ levels over time between various systems. The DQ dimensions are discussed and described in a wide variety of sources; yet, they all essentially used the same methodology, even if they frequently allude to various levels and data model components.

According to Loshin, there are two types of practical dimensions of DQ: contextual and intrinsic. Contextual dimensions consider the data element in relation to other data elements (driven by context), while intrinsic dimensions focus on the data values themselves outside of a particular data or model context.

Figure 1 shows the practical dimensions of DQ as introduced in [14].

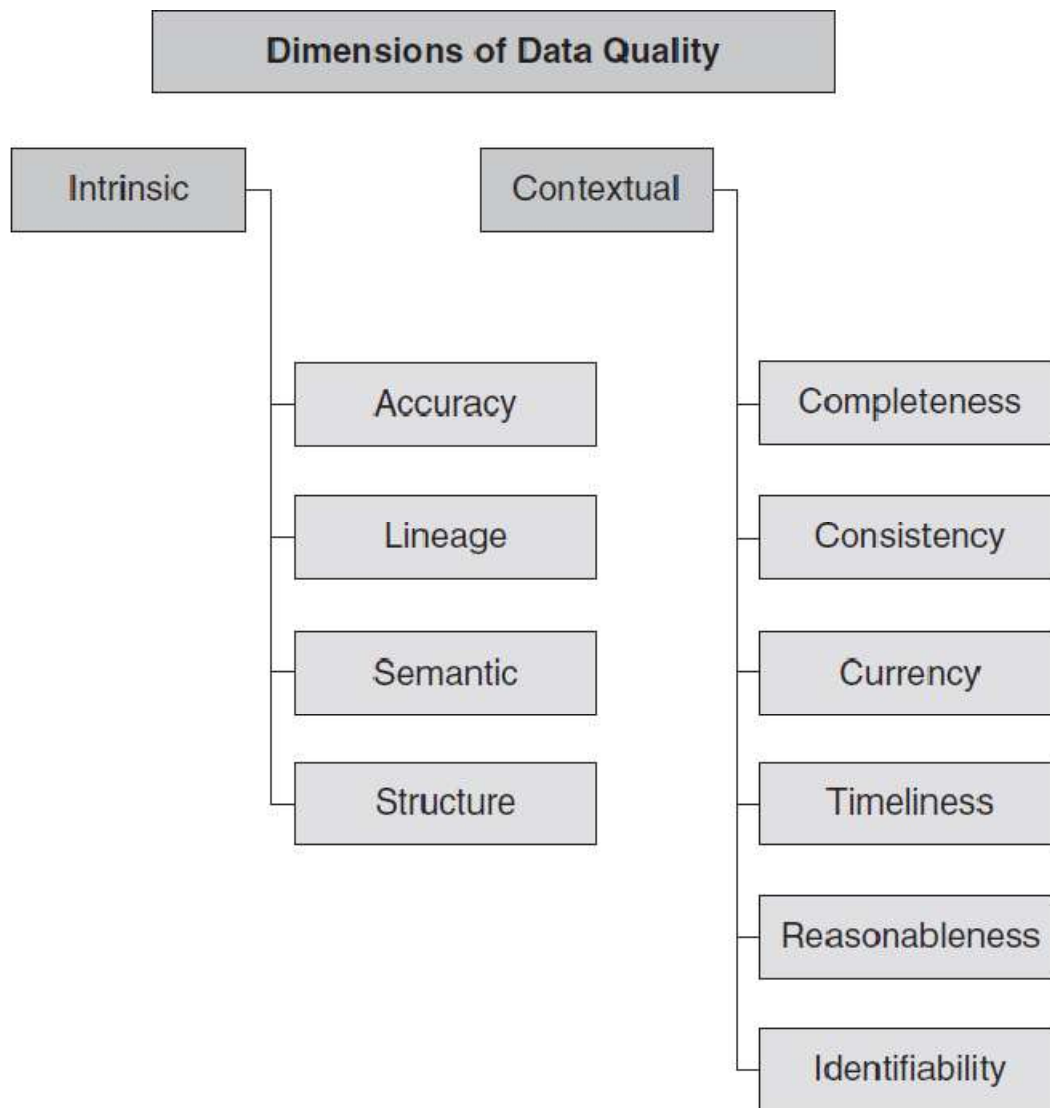


Figure 1: Dimensions of Data Quality

2.2.1. Intrinsic Dimensions

The intrinsic dimensions concentrate on the values of the data without necessarily assessing the context in which those values are found. These dimensions characterize structure, formats, meanings, and enumeration of data domains – essentially the quality of organizational metadata and how it is used [14].

The introduced intrinsic dimensions are:

- **Accuracy:** The degree to which data values agree with a recognized source of accurate information
- **Lineage:** A dimension that assesses the data's historical sources in order to determine the origin of any newly added or updated data items and, consequently, assess the data's reliability.
- **Structural consistency:** describes how similar attribute values are consistently represented within a dataset as well as across data models connected to related tables
- **Semantic consistency:** describes the extent to which related data items have consistent names and meanings. It encompasses consistency of definitions across characteristics inside a data model as well as similarly named attributes across several enterprise datasets.

2.2.2. Contextual Dimensions

The contextual dimensions provide a way for the analyst to review conformance with DQ expectations associated with how data items are related to each other [14]. The following contextual dimensions have been introduced:

- **Completeness:** describes the assumption that a dataset's assigned values for particular attributes should exist.
- **Consistency:** It is important for various data hierarchy levels, tables, databases, and applications; it also applies to externally supplied data, or, to put it another way, integrity restrictions.
- **Currency:** describes how up to date information is in relation to the world it models. Currency is a useful tool for evaluating the accuracy and up-to-datedness of information, even in the face of potential alterations or revisions that may affect time and current values.
- **Reliability:** this describes how quickly information should be available. One way to evaluate timeliness is to look at how long it takes for information to become usable after it is expected.
- **Reasonability:** This dimension covers broad claims related to expectations of values' consistency or reasonability, either within the context of current data or within a time series.

- **Identifiability:** refers to the unique naming and representation of core conceptual objects as well as the ability to link data instances containing entity data together based on identifying attribute values.

2.3. Data Quality in Insurance IS

Insurance business is one of the fastest growing businesses in the world. More than 20% of the Fortune 500 companies are insurance providers (Panwar, M. (2015)).

Almost all of the companies require essential data from the customers, which include the personal details, the employment details and in some cases, even the private medical information Or SSN [3].

Data quality is crucial for decision-making reliability. It should be complete, accurate, available, and timely, eliminating common issues like misspellings and typos, ensuring users can access necessary information on demand.

2.4. Data Quality Management Techniques

In order to manage DQ dimensions and improve it, it is important to follow a systematic process to ensure better quality within the organization and to make sure of the presence of continuous quality check [13].

Data management refers to defining the data architecture including enterprise-wide data models and data modeling standards; data administration including data dictionaries and data management processes; and data management systems including their design and operations [49].

Quality-oriented data management, or data management that concentrates on gathering, organizing, storing, processing, and presenting high-quality data, is known as data quality management.

To manage and improve DQ dimensions, it is necessary to follow a systematic approach that ensures higher quality within the organization and the existence of ongoing inspections.

As a result, the following methodologies and models have been offered from many researchers for DQ management.

2.4.1. Total DQ Management (TDQM)

It is basically an extension of Total Quality Management (TQM) framework which is using for physical product quality [13]. TDQM has been proposed to support the concept of “data as a product”, that they used the same procedures used in TQM to achieve high quality [20].

The methodology starts with the concept of an information product (IP). At this step, in order to achieve a high quality status, the IP must meet particular criteria. The information quality (IQ) measures are then created and utilized to determine the IP.

TDQM uses the *information product (IP)* approach inspired by the analogy between manufacturing product of TQM and data. Wang summarizes the purpose of TDQM as “*delivering high quality information products to information consumers*” [18]. TDQM adopts Deming’s “*Plan, Do, Check and Act*” from the TQM literature and creates its own “*Define, Measure, Analyze, and Improve*” cycle as a continuous process [21]:

1. The definition step involves identifying data quality dimensions and associated requirements.
2. The measurement step generates quality metrics. The input provided by these measures enables the comparison of actual quality with predetermined quality requirements.
3. The analysis phase determines the underlying cause of quality issues.
4. The Improvement phase focuses on enhancing quality activities.

However, when comparing information output to physical manufacturing, various concerns arise. These included the capacity to share data among users. Second, raw data may not arrive on time, and certain value assessments, such as integrity, are difficult to assign to physical manufacturing. TDQM was developed to maintain the quality of data in databases, however today's technology, notably big data, may limit its application. This is due to the wide range of data types in big data available.

The image below shows, A Summary of this methodology created by Battini et al. (2009)

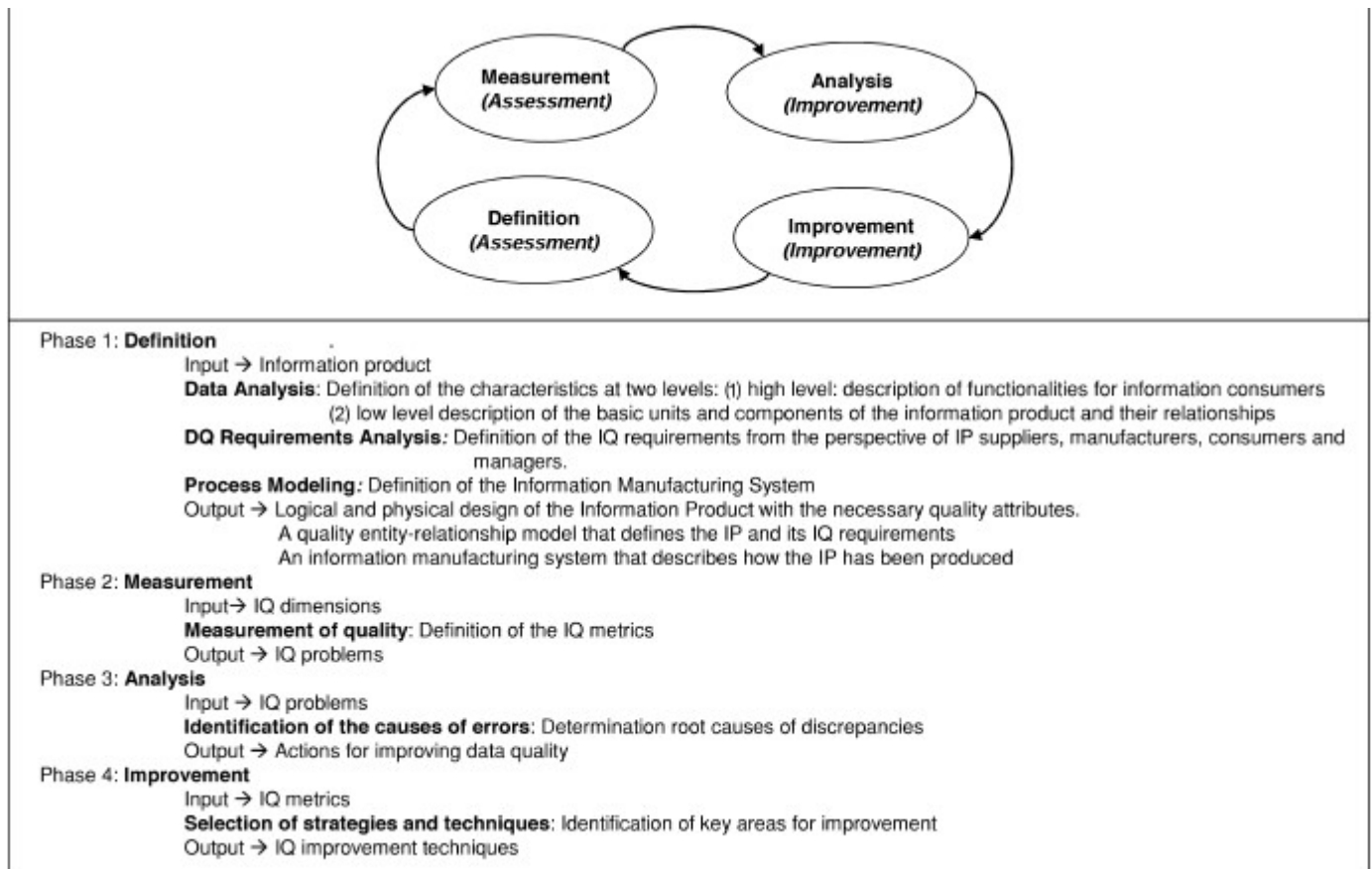


Figure 2: A Summary of the TDQM methodology created by Battini et al. (2009)

2.4.2. Information Integrity Methodology (IIM)

This methodology was introduced subsequently and emphasized the necessity to achieve information integrity by focusing on the data's foundation [12].

Information integrity took into account the organization's ability to accomplish its strategic goals. To attain high data reliability, information integrity criteria must be met. The framework includes data policies, organizational capabilities, data management, design, system, verification, interaction, and framework compliance.

2.4.3. AIM Quality model (AIMQ)

It includes the Product and Service Performance model (PSP) for IQ [19]. The Information Quality Assessment and Improvement Methodology (AIMQ) was designed to allow businesses to analyze their information quality (IQ) level [19]. Furthermore, the methodology intends to create a foundation for companies to use benchmarking methodologies to compare their IQ level

to others. It comprises the Product and Service Performance (PSP) model for IQ [22]. In this model, a questionnaire is used to assess data quality. Additional statistical analysis is then performed to categorize the problem location of data reliability.

AIMQ makes up of three components:

1. PSP/IQ model: - This is a 2-by-2 model of what IQ represents to information consumers and managers.
2. The IQA instrument:-is a questionnaire that measures IQ based on identified factors that are essential to information consumers and managers.
3. Finally, the final component includes two GAP analysis methodologies for analyzing questionnaire findings for each quadrant and its corresponding dimensions. The first technique, benchmark GAP, compares an organization's IQ to a standard. This benchmark includes the best practices of numerous organizations. The second technique, role GAP, assesses the gaps between the judgments of various participants (stakeholders) in an information production system.

2.4.5. Complete DQ Management (CDQM)

Complete Data Quality Management (CDQM) methodology aims to establish a balance in between completeness and practical feasibility of the DQ improvement process [24]. This model can handle structured, semi-structured, and unstructured data types, whereas all previous models and the majority of research focused on structured data types. CDQM proposes theoretical, empirical and intuitive approach to check the quality of the data [26] [25].

Three stages make up the methodology: state reconstruction, assessment and selection of the best improvement. In order to determine which organizational units are use which databases for which business processes, a number of matrices are built in the first phase to depict the linkages between processes, organizational units and databases. In the second phase the new target DQ levels are established with the goal of enhancing process quality

The associated costs and benefits are assessed when the new levels are determined. After that, the processes are examined to identify the most troublesome areas. In the final phase, the optimal improvement process is identified using the inputs of the previous steps and applying a cost-benefit classification to the candidate processes [26]. As mentioned before, the model is flexible

dealing with different kinds of data types, however, it does not have a clear measurement method or a way to calculate the quality dimensions which makes it difficult to apply it in the organization [13].

2.4.6. Six Sigma

Six Sigma is a management philosophy developed by Motorola that emphasizes setting extremely high objectives, collecting data, and analyzing results to a fine degree as a way to reduce defects in products and services. [3].

Sigma is a Greek letter that denotes deviation from the standard. The Six Sigma methodology seeks to quantify process flaws and methodically remove them in order to achieve near-perfect excellence.

According to Panwar, M. (2015), there are two Six Sigma processes: Six Sigma DMAIC and Six Sigma DMADV, each term derived from the major steps in the process. Six Sigma's DMADV is a process that defines measures, analyzes, improves, and controls existing processes that fall below the Six Sigma specification. Six Sigma's DMADV defines measures, analyzes, designs, and verifies new processes or products that are trying to achieve Six Sigma quality. Six Sigma processes are managed by Six Sigma Green Belts or Black Belts, who are then overseen by a Six Sigma Master Black Belt, a term coined by Motorola.

Six Sigma offers significant benefits such as up to 50% process cost reduction, cycle-time improvement, reduced waste, better understanding of customer requirements, increased satisfaction, and more reliable products and services.

According to Panwar, M. (2015), the following diagram represents the six sigma's DMADV Process.

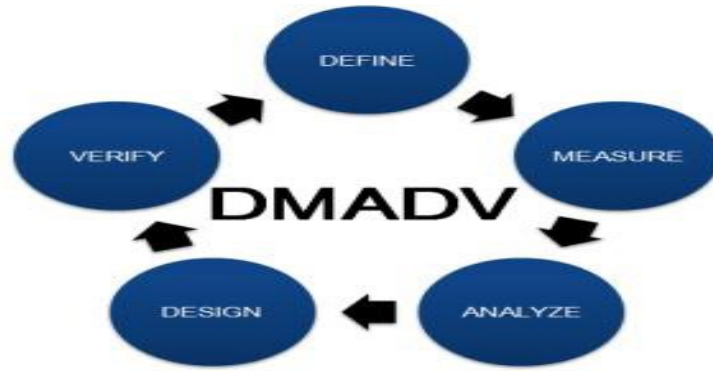


Figure 3: Six Sigma's DMADV Process

2.4.7. Master Data Management (MDM)

One of the primary reasons for the importance of MDM in Enterprise Data Governance lies in its ability to address the challenges of data inconsistency, fragmentation, and duplication [39][35].

MDM is a comprehensive approach to managing and leveraging an organization's critical data assets – its master data – in a centralized, consistent, and controlled manner [39].

Master Data Management (MDM) plays a pivotal role in this regard, serving as the foundation for efficient data governance practices within enterprises [3-table]. To maintain the accuracy, consistency, and dependability of corporate data in today's data-driven environment, effective master data management is essential.

Intends to provide a unified, standardized view of master data, including customers, products, and other information across the whole company.

2.4.8. Data Stewardship Model

Data stewardship is a term that is understood in heterogeneous ways, in recent organizational developments and efforts to build infrastructures and hire professional staff for research data management in various scientific fields in Europe, data stewardship is understood as mainly aiming at optimizing data management in line with the FAIR principles (find ability, accessibility, interoperability, reusability) for purposes of reuse in the interests of the scientific community and the public [40].

In order to preserve an effective, reliable and equitable data ecosystem, data stewardship entails making sure that the rights and responsibilities of all stakeholders are respected.

In other way, this model involves providing teams or individuals inside the company responsibility for the quality of the data. It mainly focuses on responsibility, ownership, and preventive data quality control.

Generally, all the methodology and models discussed above supports the way of managing data quality of organizations with different approaches.

The methodologies and models are customizable and extendable for the organizational business process based the nature of the system and organization's data usage way.

2.5. Review of Related Works

Some of the evaluated study works follow corrective data quality improvement procedures, while others take a holistic approach. There are further publications that are not covered in this area, such as topic-related assessments that use the Detective method.

The below study papers are briefly described, together with their recommendations and gaps which supports us for the solution design.

Vincent Buitenhuis (2023) conducted research under the title "Designing a holistic method for enhancing data quality with the use of machine learning" to improve the isolated data quality dimensions of data used for strategic decision-making using machine learning techniques. The researcher used algorithms such as Random Forest, Neural Network-based Auto Encoder, Neural Network, and KNN to solve data quality concerns such missing data, irrelevant data, incorrect data types, and outliers. However, some data quality challenges remain unresolved with the machine learning approach. Based on this, the researcher recommends that, to add features to his working prototype that address other types of data quality issues that cannot be solved via machine learning.

Creamers (2016), a study titled "Data quality improvement in a production environment" which focuses on data quality improvement in an ERP-system within a production environment. The study seeks to demonstrate whether data quality is enhanced by constructing an automation tool based on the Haug et al. (2009) paradigm, which works for the only ERP systems of a specific organization, VDL Nedcar. The researcher suggests that the technology be extended to other departments in order to improve data quality and process flow.

In thesis work titled "Improving data quality and information use of tuberculosis program in Alert hospital," Abeba (2015) developed a simple database application for managing patients' medical history and related information, primarily to solve data completeness issues, and the developed application is used primarily for registering information about patients' detail information and facilitating efficient communication among all hospital units. However other data quality issues, such as accuracy, timeliness, and consistency, still need to be addressed.

Sulanen, S. (2021).in work titled "Improving financial data quality through data governance (master's thesis)", studied how the quality of financial data can be improved by utilizing data governance to address data quality challenges. The research was conducted as a qualitative case study in which a design-based research approach was used. The findings of this research imply that data governance can be used to improve the quality of financial data in organizations because it addresses the organizational challenges that negatively affect financial data quality. However was restricted to studying a single company and data governance framework was developed explicitly for the case company. In addition to that, its effectiveness still highly depends on how well it is implemented and adopted.

VanWierst, J. J.in his thesis titled "A Process Model for organizational Data Quality Assessment" focuses on data quality assessment: “the process of obtaining measurements of data quality to determine the current state of data quality”. Based on such initial assessment, improvement plans can be made that balance data quality levels, costs, resources and capabilities across an organization or department. The research also forward one important idea that say, “majority of data quality frameworks and methodologies are either developed for a specific context, technique or problem, or they provide a generic assessment that often lacks practical guidance and is not operational zed for a specific context." his study addresses this gap between data quality assessment research and practices. The research proposed a process model for data quality assessment which is based on a synthesis. However, the model only focuses on assessment of data quality problem rather than fixing for the existing problem. one of the good things is that the researcher attempts to list out Potential improvements of the model such as the addition of a configuration guide, The addition of an extra validation loop, applying a reference model for selecting dimensions and so on.

Blomqvist, N. (2019), did his research work titled "Data quality methodologies and improvement in a data warehousing environment with financial data". The main aim of thesis is understood what is needed to successfully use a data quality methodology and give improvement suggestions with the given restrictions. And then the researcher found 13 methodologies from the literature review from which one was chosen to be used in his thesis. Among that methodology Quality Assessment Using Financial data turned out to be the best methodology with the given restrictions. The methodology uses objective and subjective assessment methods and compares their results. Based on the results, the dataset under measurement gets a quality score. If the chosen methodology doesn't provide improvement suggestions, it is recommended to revise data quality issues in the environment your data is in. Generally, the paper focuses on criteria for specific data quality problem within a given domain and implement the best amount previously selected generic methodologies, which lacks customization activity if needed.

2.5.1. Data Quality improvement in Insurance Industry

Soyer (2013) conducted a study on title "data quality management for insurance companies in Europe, focusing on the Solvency II regulation (is prudential regime for insurance and reinsurance undertakings in the EU)". The study analyzed the Solvency II regulation's directive's requirements using Requirements Analysis techniques and then transformed them into system specifications, and explained the necessary implementation tasks for each specification, which is used to enhance data quality for Solvency II regulation. However, the study's findings are limited to the regulation of Solvency II and should be expanded to include other data quality difficulties in other fields.

Models developed by Duma, M. S. (2012) under the study titled "Improving Classification Performance in Missing Insurance Data" aid the supervised learning approach maintain resilience and accuracy when used with insurance datasets. His efforts, however, did not address other aspects of data quality; instead, they were limited to missing data.

In 2023, Mbithi, E. M. did study work titled "Implications of data quality dimensions for internal audit performance in Kenyan insurance companies." The following particular goals were to be the primary focus of the study: The first step in analyzing how data security control affects insurance companies' internal audit performance is to Secondly, to investigate the impact of data completeness on insurance companies' internal audit performance. Lastly, in order to investigate

how reliability of data affects the internal audit performance of insurance businesses in Nairobi, Kenya, the fourth goal is to ascertain how time-related factors affect the internal audit performance of insurance businesses in Nairobi, Kenya and how audit professionalism moderates the relationship between data quality and internal audit performance. The benefits are highlighted by the outcomes analysis. After all the study attempts to find out the effect of some data quality dimensions (completeness, reliability and timelines) only on internal audit performance of insurance companies in Kenya, however other specific data quality dimension has not been identified as initiative for enhancing decision-making and trustworthy reporting.

Chapter 3: Research Methodology

For the achievement of the objective of this thesis, this research uses Design Science Research Methodology (DSRM) because it aligns well with the practical challenges faced by Awash Insurance S.C. in improving data quality within its information systems, offering a systematic and iterative framework for addressing these challenges and delivering tangible benefits to the organization. Design Science Research (DSR) is a problem-solving paradigm that seeks to enhance human knowledge via the creation of innovative artifacts [27].

Design Science Research seeks to enhance human knowledge with the creation of innovative artifacts and the generation of design knowledge (DK) via innovative solutions to real-world problems [27]. These artifacts aim to improve and solve problems in their respective contexts.

The main purpose of this research is to enhance data quality concerns within Awash Insurance's information system by identifying particular data quality problems and then proposing appropriate solutions using design science research problem solving procedures.

DSR results in Information System (IS) have been shown to create significant economic and societal impact (Gregor & Hevner 2013, vom Brocke et al. 2013).

An overview of the DSRM process described when applied to the creation of information systems in a series of activities is provided by Peffers et al. (2007).

Figure 3.1 shows the process of the methodology of Peffers et al. (2007).

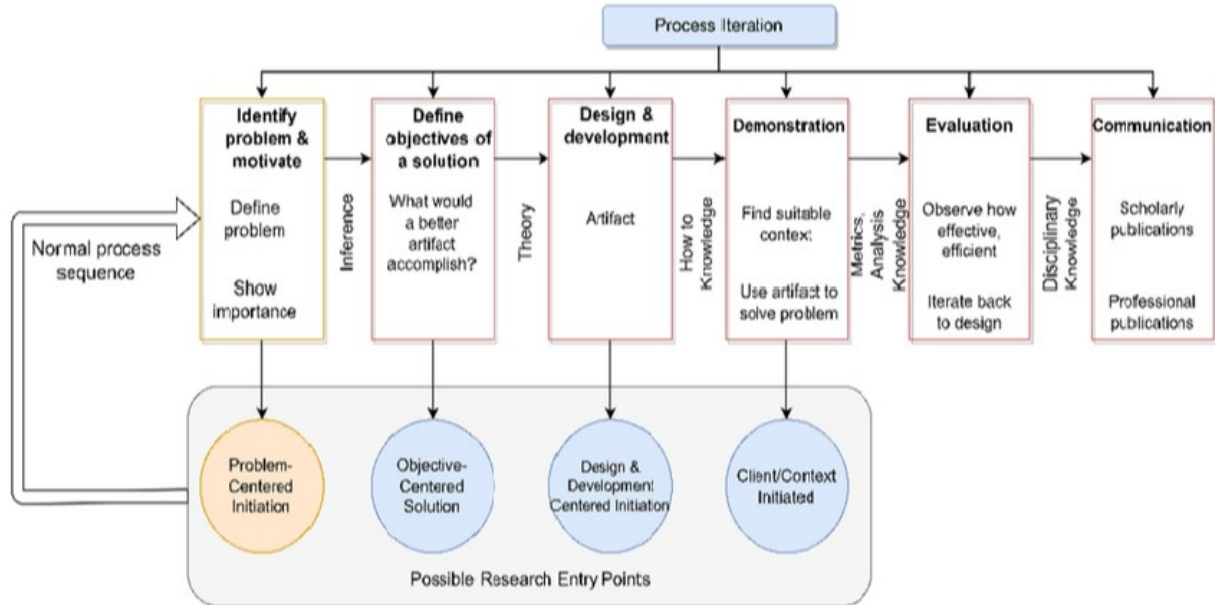


Figure 4: Design Science Research's Process model (Peppers et al., 2007)

By integrating Design Science research methodology with appropriate data collection techniques, the thesis can ensure a systematic and robust approach to improving data quality for Awash Insurance S.C.'s information system.

The following section describes the followed paper's research methodology (peffers et al., 2007).

3.1. Description of study's research process methodology

3.1.1. Problem identification and Definition

The stage where the identification of problem to which the creation of an artifact or solution is defined, that creates an understanding on the problem and justifying of the research. As stated in Chapter 1, Section 1.2 provides a detailed presentation of the research problems statement in order to decide on and implement a suitable approach for data quality enhancement in Awash Insurances' Information System.

The identification and definition of problem serves as motivation for the creation of an artifact (in this case, Insurance Data quality Improvement process model) that may effectively give a solution.

Before beginning this research, a detailed systematic literature review was conducted to assess and explain organizational data quality analysis and enhancement, as well as the current state of knowledge in the subject.

The majority of data quality frameworks and methodologies are either developed for a specific context, technique or problem, or they provide a generic assessment method (i.e. regardless of context or application) that often lacks practical guidance and is not operational for a specific context or business needs (van Wierst, 2019).

In this study, the solution to be provided, the Insurance DQ Improvement Process model, assesses and improves data quality issues of the company through the use of an appropriate systematic process.

At this stage, the researcher reviewed literatures to obtain data for this approach and then build the scientific background.

The researcher first noted the foundations of a data-driven Organization, mostly insurance companies, and then studied the usage of data in such an environment through reviewing data and information systems and their various components. This supported us to address the first research questions.

Furthermore, literature has been reviewed to explore the practical consequences of data quality. This was accomplished by determining which aspects of data quality should be examined and which data quality concerns are commonly faced. This let us to respond to the second research question.

Lastly, based on previous results, we explored which solution methods or models are commonly used to address particular data quality difficulties within the stated scope. Meaning, data quality issues are generally seen in the insurance industry, particularly Awash Insurance.

3.1.2. Definition of the objectives for a solution

This phase involves setting research goal and objectives based on the problem statement. This involves establishing limits and scope of what is possible and practically achievable. Sections 1.3, 1.4 and 1.7 of the thesis discuss this activity.

The literature review provides background for scientific knowledge and technological potential.

The study's defined specific goals, which are mentioned in section 1.4 of this paper, designed to directly fulfill and address the research questions that were taken from the problem statement of the study.

3.1.3. Design and Development

Following a clear definition of the problem and the objectives that a solution has to achieve, the next step is to create the artifact, in this case an Insurance Data quality Improvement process model. The artifact's functionality and layout are determined prior to development.

According to Peffers et al. (2007), moving from objectives to design and development needs the incorporation of theoretical knowledge into a solution.

Prior to developing the model, the following knowledge is required: Identification of the relationship between Insurance and data, data quality concept and Data quality dimensions, Generic management methodologies and as well as domain-specific data quality assessment and improvement process ,which are acquired from detail literature review as stated in chapter 2. In addition to this, semi structured interview is conducted(detail description of data collection will be discussed in section 3.2) with Awash Insurance experts so as to understand the companies' business process that uses Integrated Information system and to know factors that affect data quality problem in the company, This let us to respond to the first research-question. Additionally, we formally requested raw data with data quality issues for this paper's demonstration phase, which will show how beneficial the model is. This allowed us to answer the second research question.

Lastly, Model called Insurance data quality improvement process model is provided as a new artifact in the area of study, which is appropriate for improving data quality issue for Insurance Information system, particularly which will be applicable to Awash Insurance company's Information system, in which the final research question of this paper will be answered.

3.1.4. Demonstration

This stage involves a proof of the artifact or solution generated from the research, which is a necessary part to show its functionality and effectiveness in a real-world setting. As this thesis research attempts to give a solution to doing data quality improvement in the field, its

demonstration should be in the field as well, with simulation of raw data with DQ problem to the model (detail will be shown in chapter 5).

As a result, a case study was chosen as the approach for demonstrating how to use the solution model.

3.1.5. Evaluation

It is an activity of observation and measuring the degree to which the artifact provides a solution to the defined problem. To determine this, the previously specified solution objectives will be assessed using the demonstration's observations and findings. On the basis of this assessment, the research either returns to the design phase to improve efficiency or proposes prospective changes for future research or projects with the help of relevant stakeholder involvement.

The method of evaluating the artifact (model) will be discussed in Evaluation chapter of this paper.

3.1.6. Communication

The final steps that involves communicating every details of the process end to end to all relevant audiences. The target audience might comprise data quality researchers, professionals employed in the field, and Academicians.

The following section discusses the approach of data collection, analysis and finding for the support this iterative problem solving process.

3.2. Data Collection Methods

In this section the method of data collection techniques will be discussed.

3.2.1. Data Collection

The data collection method for this thesis, involves qualitative approaches, depending on the specific objectives of the study. Sources of data collection methods that could be used are discussed as follows:

1. Primary Data source

As a primary data source, Interview and document analysis will be used, so as to gain valuable insight for study work.

Interviews: Semi-structured interviews were conducted with 11 participants among key stakeholders within the organization as per the sampling strategy (discussed in the section 3.2.2). Those interviewed were taken from IT personnel's, data Analysts, and insurance Customers, that could provide more in-depth qualitative insights into the challenges and opportunities for improving data quality. Interviews allow for open-ended questions and probing exploring issues in greater detail as attached in appendix I.

Document Analysis: reports, system manuals and policies related to data management and quality assurance within Awash Insurance S.C was reviewed and analyzed with Content Analysis approach that could provide additional context and insights into existing practices and challenges. Some of reviewed documents are attached in Appendix II.

Raw Data: In addition to the above sources, raw data generated from the company's system software for model simulation during demonstration activity has been used.

2. Secondary Data Source

As a secondary data source, systematic literature review process was conducted. Based on the aim of the research, review of the concepts related to the subject matter is important that focus on collecting and analyzing existing studies, theories, methods and models or frameworks that are relevant to both the problem domain and the design of the artifact or solution to be offered.

The researcher assesses these criteria in a random manner as mentioned below.

Criteria for Inclusion:

- The literature should be directly related to improving data quality in insurance information systems or has or similar contexts, ensuring it aligns with the thesis topic.
- To ensure the reliability and correctness of the information incorporated in the review, the literatures should be obtained from recognized academic search engines such as Google Scholar, Research Gate, and so on.
- The methodology or procedure given must be validated through experimentation or case study research to demonstrate its verified value in increasing the quality of data.

Criteria for Exclusion:

- Literature works published before fifteen years ago should be excluded from the literature review section.
- The literature that is not available in researcher language(s) or relevant to the target audience must be excluded.

3.2.2. Sampling Strategy

The suggested approach for designing a sampling strategy for semi-structured interviews with key stakeholders within Awash Insurance S.C. for improving data quality in their information system involves several considerations.

Identifying Key Stakeholders (Target Population)

Key stakeholders within Awash Insurance S.C. who have direct involvement or expertise in the data quality of their information system were identified. This includes System users, IT professionals, data analysts, insurance customers, managers from underwriting and claims departments.

Determining of appropriate sample type

Purposive Sampling was employed to select participants with diverse perspectives on data quality issues and improvement strategies, ensuring representation from various departments and levels for a holistic view.

Determine Sample Size

Using a thoughtful approach that considers the research goals, population, and practical constraints, making a manageable sample size of 10-15 participants is appropriate for this case. So based on that, one participant from each of two departments (Claim and Underwriting), three participants among IT professionals, data analysts and insurance customers, which is a total 11 participant are involved in interview activities.

3.2.3. Data Analysis Method and Finding

Interview Analysis and Result

Based on the research objectives, interview activity aims to explore several areas of data quality, such as current challenges faced by Awash Insurance S.C., potential causes of these challenges, and the impact of data quality issues on the organization's efficiency and effectiveness.

Furthermore, explores the roles of stakeholders, including employees, customers, and external partners, in ensuring data quality within Awash Insurance S.C. It also examines the resources and support required by the organization to address data quality issues effectively.

Additionally, the study identifies key priorities and areas of focus for improving data quality within the insurance information system at Awash Insurance S.C. It considers how the effectiveness of proposed solutions for data quality problems will be evaluated and invites any additional comments or remarks on the subject. Interview Question checklist is attached in *Appendix I*.

Thematic Analysis approach is employed for details analysis of respondent's answer though hearing the interview recordings several times, the following finding is drawn.

Among respondent's, IT professionals are more familiar with the concept of data quality than others, whereas regarding the impact of data quality problem on company's performance, Managers from Underwriting and claim section forwards important point.

Regarding the type of data quality faces by AIC; the respondents answer ranges from data accuracy, completeness, consistency, timeliness, to relevance. In addition to that, factors like data entry errors, system limitations, or inadequate data governance practices are identified as the main causes of DQ problem in AIC' Information system.

Regarding, assessment of how these data quality issues affect the overall efficiency and effectiveness of insurance processes within Awash Insurance S.C, it has high significant on Underwriting premium collection and claim processes, that leads to delays, errors, or inefficiencies in core business operations of the company.

And also as solution, resources such as technology, training, or personnel, needed to address data quality issues effectively within Awash Insurance S.C.'s information system is highlighted by respondent.

For that of determination of key focus areas or priorities for improving data quality within the insurance information system at Awash Insurance S.C., Core business process of the company like Underwriting and claims are identified as key focus area.

Regarding evaluation of the effectiveness and impact on business outcomes of solution to be proposed, implementing to the real word data and feedback from relevant stake holder is critical.

Document Analysis and Result

With Content Analysis approach, it is attempted to analyze company's (AIC) documents like system manual, website content and annual financial statement report that helps to identify the company's core and supportive business process as well as, the impact of data quality on its decision making process.

The finding from analysis result revealed that, Underwriting and claims are among core business process of the company which faces most of data quality challenges and the data used for these processes highly affects the company's decision making process.

3.2.4. Ethical Considerations

To ensure data privacy and confidentiality, after receiving approval Letter from Addis Ababa University's School of Information Science, the study work was carried out. In addition to this, the purpose of the study for collecting relevant data is clearly informed to company's (AIC) participant employee with the measure taken to de-identify the data to protect the identities of individuals involved. In addition to that, data used for model simulation purpose will be encrypted for security purposes as it will be shown in demonstration phase. By addressing the above ethical considerations, it is attempted to conduct responsible and ethical research, especially when dealing with sensitive data of the company (AIC).

Chapter 4: Design and Development.

This chapter focuses on model design and methods for improving the quality of data for the Insurance Information System, with a focus on Awash Insurance Company.

A literature review is carried out to gather the information needed for the process model's design. Primarily we will start by explaining the considerations that contributed to the design. To do this, list of every requirement and stakeholders have been identified before the design.

Other than finding research directly from the databases, contributions are found by checking relevant references to other work as well (Batini et al. 2009). Appendix III shows the final set of search words used in the literature review.

In this study, some of known Generic Data quality Management methodology and a few domain specific models designed by other researchers has been reviewed and analyzed. These models are derived and customized from generic methodologies mentioned in literature section.

Lastly, considering the listed requirements and Insurance business process (particularly for Awash Insurance Company) as well as the gap of the domain specific model designed by other researchers, the customized Data quality improvement process model to be designed and developed as solution artifact.

4.1. Requirements Specification

Prior to a design, we have identified a list of specifications of requirements suitable for our Data quality improvement process model. We start by discussing the process of determining and collecting the requirements. Following that, we classify and enumerate those requirements as functional or non-functional. Finally, we explore and describe the various main actors (stakeholders).

4.1.1. Approach of requirements Specification.

A literature review is carried out to collect the knowledge essential for the process model's design. Following the finding of interview and document analysis, a review and analysis of literature on information systems, insurance data, data quality, data quality dimensions, and data quality management approaches (both generic and domain-specific Models developed by other researchers), requirements were determined and specified.

The solution (in this case, Insurance Data Quality Improvement process model) attempts to enhance the quality of data in the company's insurance information system, influencing strategic decisions.

4.1.2. Functional Requirement.

Robertson (2014) defines functional requirements as follows: “*Functional requirements specify what the product must do — the actions it must perform to satisfy the fundamental reasons for its existence.*”[30]. Identifying functional requirements for the Insurance Data Quality Improvement Process will drive the tool's (in this case Insurance Data Quality Improvement Process Model) functionality and quality of the desired results.

Selection of DQ improvement Process model requirements depends on the conclusions drawn from section 3.1.2.

List of Functional Requirements for Insurance Data quality Improvement Process Model

- The model should be capable of improving the quality of data used for insurance company's (Particularly for Awash Insurance Company) business processes and strategic decision-making processes.
- It has to be capable of addressing inaccuracy issues of data values for a given dataset.
- It has to be capable of resolving inconsistencies issues in a dataset's values.
- It should be able to eliminate redundancy issue within a given dataset.
- It has to be capable of detecting and addressing incompleteness issues of data values for a given dataset.
- It must follow the right approach or procedures depending on the type of data quality issue that has to be resolved.

4.1.3. Non Functional Requirement.

Non-functional requirements (NFRs) establish criteria that describe a system's behavior rather than its function.

According to Robertson (2014), Non-functional requirements are defined as “properties, or qualities, that the product must have if it is to be acceptable to its owner and operator”.

We have identified non-functional requirements based on the solution system's characteristics [2].

List of Non Functional Requirements for Insurance Data quality Improvement Process Model

- **Performance**
 - Describes how well the system operates under specific parameters, such as response time, throughput, scalability, and resource utilization. In this case, the model performs data quality enhancement for identified and specific DQ problem following the process stated for the improvement. Its effectiveness is also measurable in terms of enhanced dataset's quality.
- **Usability**
 - Defines the system's usability and user experience, including issues like user interface design, accessibility, and user documentation. For this case, the proposed model effectiveness is demonstrated and evaluated with relevant stakeholders.
- **Scalability**
 - Refers to the system's ability to manage increased workloads and scale up or down as needed without significantly impacting performance or functionality.
 - The models can be used to solve similar problems with data quality in other company units.
- **Maintainability**
 - Specifies how readily the system can be improved, updated, and maintained over time, taking into account code readability, modularity, and documentation quality. The proposed model would be designed and developed to be easily maintainable, updatable and improvable.

4.1.4. Stakeholder Identification.

There are several relevant factors involved in our suggested method for improving the data quality of Insurance Information system. A stakeholder can be defined as a person with interest in the use and outcomes of this method [41]. Identified stakeholder is listed as follows:

Insurance Company's (AIC) IT professional:

Due to the dependence on accurate information for underwriting, risk assessment, and claims processing, insurance firms' IT professionals are the main stakeholders. Improved data quality benefits them indirectly (especially professionals like Application Support Officers). Because it reduces fraudulent activity, minimizes errors, enhances the efficiency and effectiveness of business process and decision-making.

Data Analyst:

The process of making decisions based on data is assisted by data analysts. They conduct insightful business analyses and evaluate the data that is available within a company. To ensure high-quality insights, data quality is an essential requirement.

Policy Holders:

Because data quality directly impacts policyholders' insurance coverage, premiums, and claims experience, they have a stake in it. Accurate data guarantees that policies correspond to their exact risk profiles and that claims are resolved appropriately and properly.

System Users:

System Users are individuals who have been allowed to use a company's system software for internal business processes. The role of a system user has a direct impact on the quality of data utilized for the company's system software since they have access to enter new data into the system and manipulate data that has already been entered.

4.2. Description of Literature Review process

An important stage that guides the design and development of new artifacts, systems, or solutions is the literature review process.

As mentioned in chapter 1 of this paper, the study aims to improve data quality problems that exist within Insurance Information System particularly for Awash Insurance Company. Hence it is important to focus on collecting and analyzing existing studies, theories, methods and models or frameworks that are relevant to both the problem domain and the design of the proposed artifact or solution. The following section discusses the way of selecting, searching and analyzing of the literatures in detail.

4.2.1. Approach of Literature Selection for Solution Design

As a selection criteria, Our literature review included an extensive variety of areas, including information systems, insurance data, data quality, data quality dimensions, and data quality management Methodologies (including both generic and domain-specific Models developed by other researchers).

Apart from the literature review, we have obtained and incorporated the system reports of the company (which contains data quality issues) as a source of data for the model demonstration phase. Additionally, we conducted a semi-structured interview with relevant stakeholders (Experts) from AIC to gather and analyze additional input.

For searching Strategy, The context of this study and the research questions define a set of search terms.

Other than finding research directly from the databases, contributions are found by checking relevant references to other work as well (for example provides Batini et al. (2009) many relevant references).

Appendix III, shows complete list of search terms used will be shown by the search terms used for the literature review.

Additionally, the researcher uses a set of inclusion and exclusion criteria to assess the study contributions for this literature review and the problems that still need to be answered as mentioned in section 3.2.1 of this paper.

4.2.2. Summary of Interview result

As mentioned in section 3.2.3 of this paper, Data quality issues at Awash Insurance S.C. include accuracy, completeness, consistency, timeliness, and relevance. Factors like data entry errors, system limitations, and inadequate data governance practices are identified as main causes. To address these issues, resources like technology, training, and personnel are needed. Key focus areas include underwriting and claims.

4.2.3. Summary of Document review analysis.

The analysis of company documents like system manuals, website content, and annual financial statements helps to identify the company's core business process as underwriting and claims and the impact of data quality on decision-making.

4.3. Analysis of Existing DQ Management methodology and Model customization

This section aims to provide an overview and analysis of the existing generic data quality management approaches and domain-specific models. Additionally, it will discuss the necessary procedural way of customizing them to fit our organization's specific needs.

4.3.1. Detail description and Analysis of DQ Management methodology.

As mentioned in section 2.5, the detail description of common generic data quality management methodologies was presented. Here, we extract information on the fundamental concepts, schemes, phases and best practices of these methodologies and models for facilitating the analysis and customization attempts, and evaluating it from the perspective of the requirements for the insurance data quality improvement process.

Table 1 : Data Quality Management methodologies analysis

Methodology	Overview	Key Features and Focus	Application in Insurance Data	Level of Applicability
Total DQ Management (TDQM)	Emphasizes the use of technology, procedures, and people to manage data quality comprehensively throughout the company.	Focuses on integrating DQM into corporate culture, data governance, and continuous improvement.	Preserves the accuracy of policy data, claims data, and client information for risk assessment and compliance.	Very High
Information Integrity Methodology (IIM)	Focuses on implementing different procedures and controls to ensure the reliability and integrity of information assets.	Provides a strong emphasis on data timeliness, consistency, accuracy, and completeness.	Ensures consistency in the details of the policy, the computation of the premium, and the handling of claims.	Medium

AIM Quality Model (AIMQ)	Offers a framework for evaluating and enhancing the quality of data using six criteria, such as accuracy, consistency, integrity, etc.	Ensures that data satisfies regulatory criteria and aims to match data quality with business objectives.	Supports in preventing underwriting errors, detecting fraud, and preserving accurate policyholder information.	High
Complete DQ Management (CDQM)	Comprises monitoring, remediation, data cleansing, profiling, and monitoring as well as all other DQM features.	Attempts to preserve a high standard of data quality across every stage of the data lifecycle.	Maintains accuracy and consistency in the processing of claims, regulatory reporting, and policy management.	Medium
Six Sigma	Identifies and removes errors in the process, especially those related to data quality, using statistical techniques.	Aims to enhance overall process performance by minimizing variability and errors.	Supports insurers in eliminating errors and inefficiencies in customer service, claims processing, and underwriting.	Very High
Master Data Management (MDM)	Intends to provide a unified, standardized view of master data, including customers, products, and other information across the whole company.	Guarantees the accuracy, consistency, and synchronization of important data across systems.	Assists insurers in effectively managing policyholder data, product details, and distribution connections.	Medium
Data Stewardship	Involves providing teams or individuals inside the company responsibility for the quality of the data.	Puts an intense focus on responsibility, ownership, and preventive data quality control.	Allows insurers to establish different roles and duties for upholding standards for data quality.	Low

4.3.2. Conclusion of the Analysis of the Methodologies and model.

The purpose of examining the Existing Data quality Management Methodologies and models from the perspective of Insurance Information system and its data usage is, for implementing robust processes to ensure the accuracy, reliability, and compliance of insurance data. Insurance firms may reduce risks, enhance customer satisfaction, and make better decisions by using proven data quality improvement models and methods.

The methodologies and models are analyzed using four parameters: overview of the Methodology or model, Key Features and Focus, application in Insurance Data and its Level of Applicability to Insurance Information system and its data usage. This method of analysis helps to prefer the best and more appropriate method or model, that will be used as a basis for the solution(In this case Insurance Data Quality Improvement Process model) that will going to be provided in the paper.

The finding of analysis activity revealed that, almost all of the analyzed methods and models are useful for improving data quality in insurance information system and its data usage; however the degree or level of its applicability is different. The parameters used to analyze the methods and models support to identify and rate its applicability level.

Among the analyzed methods and models, Total DQ Management (TDQM) and Six Sigma are those rated as highly applicable for solving data quality issues with Insurance firm business process, Where as it lacks some features and procedures such as involvement of stakeholder, unfit model component (in case of Six Sigma model) and in case of Total DQ Management, validation phase and Clear definition of processes were missed, which needs a little bit customization and creation of appropriate model for fixing the research's defined problem.

4.4. Design of the customized Model

In this section, we provide a solution to primarily defined data quality issues for insurance businesses, specifically for Awash Insurance Company, through analyzing and customizing existing data quality management and Improvement methodologies and models.

As mentioned in section 4.3.2, the final finding from analysis process were driven and used as a basis to develop robust solution (In this case, Insurance Data Quality Improvement Process model).

The design and development of Insurance Data Quality Improvement Process model will be done by customizing and incorporating two different existing methodologies that have been rated as highly applicable to insurance organizations. This involves synthesizing the strengths of each approach while addressing the specific requirements and context of the insurance industry's business process, particularly for Awash Insurance Company.

The synthesizing of the approaches' or methodologies' strengths and drawbacks from Insurance industries perspective is based on the following concepts:

- **Understanding of the Methodologies:**

- **Six Sigma** is a process improvement methodology that employs data-driven methodologies and statistical tools to reduce faults and variation. In this case, it helps the insurance companies in eliminating errors and enhancing data quality in customer service, claims processing, and underwriting, by using one of the Six Sigma process which is DMADV (defines, measures, analyzes, designs, and verifies).
- **Total DQ Management (TDQM)** focuses on the use of technology, procedures and people to manage data quality holistically across the organization. In this case, the model component has a feature that includes standards and protocols, selection of appropriate techniques and relevant stakeholder for Insurance data quality Improvement.

- **Identification of Common Components:**

- Both approaches focus on the relevance of data-driven decision-making.
- They set priorities for client satisfaction and process efficiency.
- They promote a systematic approach to quality management.
- They follow "*Define, Measure and Analyze*" process phases in its continuous improvement process.

- **Integration of Key features of procedures**

- **Define:** Begin by defining data quality objectives, aligning them with company objectives, and identifying relevant stakeholders.

- **Measure:** Assess the current level of data quality using Six Sigma's data measurement tools. TDQM's focus on customer satisfaction can help influence the selection of appropriate quality indicators.
- **Analyze:** Identify the root causes of data quality concerns using Six Sigma statistical tools. The emphasis on employee involvement in TDQM allows cross-functional teams to analyze data quality issues.
- **Design:** Create improvement strategies using Six Sigma's DMADV (Define, Measure, Analyze, Design, and Verify) method. TDQM's focus on continuous improvement can guide iterative enhancements to data quality processes.
- **Verify:** to verify and evaluate the new approach and instruction that will be used within the company to improve data quality.

- **Adaptation to Insurance Organizations Context**

The implementation of customized Insurance Data Quality Improvement Process model adopted from Six Sigma and Total Data Quality Management (TDQM) approaches to improve data quality for insurance companies requires, a focused approach to assure reliable, accurate, and up-to-date data, which has significant effect to enhance decision-making, improve regulatory compliance, and efficient business process.

To adapt those approaches to insurance organizations, the following factors must be considered.

- Organization's specific characteristics, culture, and resources needs to be considered. In this case, the companies integrated Insurance Information system along companies system usage guideline was taken as input to be considered.
- Customization of the process model to align with existing workflows, roles, and responsibilities. In this case, Awash Insurance Company's system authorization workflow and users role based on their respective job position is considered. And finally.
- Feedback from relevant stakeholders for the support of successful implementation of the model has been taken as the main input of study work.

4.4.1. Proposed Insurance Data Quality Improvement Process Model

The Model (IDQIP) chosen for the thesis due to its structured, industry-specific, and comprehensive nature, and its potential to achieve research objectives.

The proposed model offers a comprehensive approach for enhancing data quality in the insurance industry particularly design for Awash Insurance Company's Information System, by integrating Six Sigma's DMADV process methodology with Total Data Quality Management principles. Each step comprises specific activities aimed at resolving data quality issues, developing effective solutions, validating their effectiveness, and ensuring continuous improvement over time.

Both methodologies (Six Sigma's DMADV and TDQM) are briefly reviewed and summarized as follows:

Six Sigma is a management philosophy developed by Motorola that emphasizes setting extremely high objectives, collecting data, and analyzing results to a fine degree as a way to reduce defects in products and services [3].

Potential benefits of Six Sigma

Six Sigma proponents claim that its benefits include up to 50% process cost reduction, cycle-time improvement, less waste of materials, a better understanding of customer requirements, increased customer satisfaction, and more reliable products and services [3].

The **TDQM** process was originally designed for data warehouse projects on ensuring that quality is kept at a certain or desired level [2]. It was first introduced by Richard Wang in 1998 in his work called a product perspective on total data quality management (Batini & Scannapieca, 2006).

The methodology consists of four main processes: Define Measure, Analyze and Improve [2].

Now the final solution or artifact to be proposed for our case is designed in consideration with the above mentioned preconditions.

The following diagram shows the proposed customized Model named as Insurance Data Quality Improvement Process Model.

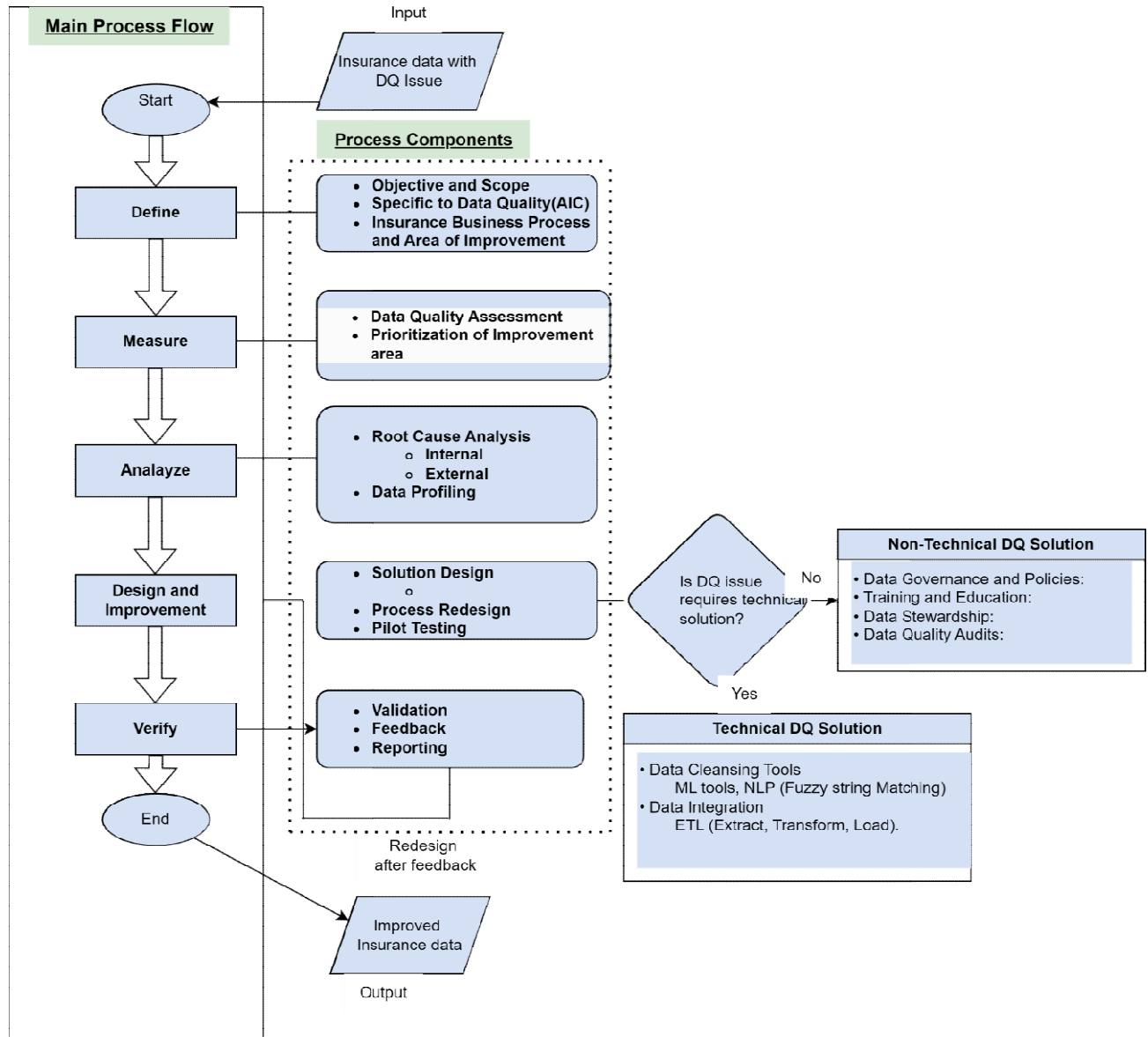


Figure 5: Proposed Insurance Data Quality Improvement Process Model

4.4.2. Description of the Proposed Data Quality Improvement Process Model.

Here, we briefly explain the process and activities of each phases of the model. This shows the plan for the execution of the model.

1. Start

The model process starts at this phase. At this stage of the model, the identified organizations information system's data that has data quality problem will be taken as input for the improvement process.

In order to ensure data quality and approval, a comprehensive process that includes collecting, converting, and publishing data must be followed. The integration of Six Sigma and TDQM principles offers a systematic method for data collection, analysis, and control, thereby enhancing data quality.

2. Define

Insurance companies primarily rely on customer inputs during customer information, proposal and policy detail registration processes, and the input will be obtained from various sources like government offices, customer service and others sources. In this phase of the model, it is essential to define, specific data quality problems encountered during the data acquisitions process. In addition to that, defining the business goal of the organization, the scope and area of the DQ improvement process task, is important because, it is used as a basis in the process of assessing and improving of specific data quality problem of the identified company's Information system.

3. Measure

In this phase, it is very important to identify the type of data that the Insurance Company uses for its business process. And also to what extent those data's are free of errors and meet data quality standard of company's data management legacy systems. Lastly, prioritizing area of the data quality improvement of the company's information system data, based on the impact it has on the organizations business outcomes.

4. Analyze

The information gained from the measure phase can be used further assessment and to address the data error or defects' root causes. This will be performed through gathering relevant information from appropriate sources.

According to Panwar.M (2015), the cause of data quality in Insurance Companies is stated as follows:

“The specific technical problems include missing data, misfielded attributes, duplicate records and broken data models to name just a few. But rather than merely patching up bad data, most experts agree that the best strategy for fighting data quality issues is to understand the root causes and put new processes in place to prevent them. Below are some common causes.

External Causes

- (1) Data errors caused by data coming from outside the organization.*
- (2) Data errors caused by external customer*
- (3) Data errors caused by manual mistakes by customers.*
- (4) Data error due to inconsistent data received from group customers*

Internal Causes

- (1) Data errors caused by human error in customer care*
- (2) Data error due to incorrect codes*
- (3) Data error due to complex databases OR older data.*
- (4) Data error due to change in standards.*
- (5) Data error due to merger/acquisition”*

In our case, the cause of data quality problems of the Insurance Company (in this case AIC), will be obtained from detail analysis and understanding of Insurance business processes and the way the it is managed by a company's (AIC) Information system, In addition to this, finding from primary data collection (interview with relevant stakeholder and document review activity) will be used as additional input. After detail analysis of the root cause of the problem, relevant summaries of data will be created, that is what we call it Data Profiling. Data Profiling is to understand the dataset at hand and its metadata [52].

5. Design and Improvement.

This phase of the model (IDQIPM) is a core to DQ improvement process, in which the effective solution to respective identified data quality problem will going to designed and implemented. In addition to this, the designed solution will also be redesigned, based the result after implementation and feedback from relevant stakeholders, if any.

The solution design for data quality improvement process can indeed be broadly classified into technical and non-technical categories, because not every data quality problems are solved either only technically or non-technically, so based on the type the data quality issue encountered, the effective solution under those categorized will be selected and applied.

The activities of preferring and implementing the effective solution will be performed by relevant stakeholder such as Data quality improvement experts, IT professionals and system users from respective company's unit.

The various solutions options proposed in design and improvement stage the model (IDQIPM), based on finding of previous data analysis, and which was categorized as technical and non-technical will briefly discussed below.

Non-technical Data Quality Improvement Solutions

Based on the type of specific data quality issue, the appropriate solution will be selected and implemented, and then the effectiveness of the solution will also be verified.

Standard and protocol: For this solution component of design and development phase, the researcher recommends, Panwar M. (2015)'s idea as stated below:

“When data quality problems are encountered importing data into the data warehouse, there are four viable actions that can be taken. We can:

- *Reject the error*
- *Accept the error*
- *Correct the error*
- *Apply a default value for the erroneous data*

When accuracy is more important than completeness, it may be appropriate to reject the error.

When the data is known to be erroneous, but it is within the tolerance level, then it is appropriate to accept the error. When the correct value can be determined, then the error can be corrected.

Lastly, when the correct value cannot be determined and completeness is very important, then a default value can be substituted for the erroneous data.” [3].

Training and Education:

In order to prevent human error in data input, customer service representatives and other individuals involved in data gathering must receive training, which means educating staff on data quality best practices and the importance of accurate data entry. In addition to this, implementing strategies to support the adoption of data quality improvements across the organization is very important.

Data Stewardship: As explained in chapter 3 of this paper, in order to preserve an effective, reliable and equitable data ecosystem, data stewardship entails making sure that the rights and responsibilities of all stakeholders are respected.

In other way, this model involves providing teams or individuals inside the company responsibility for the quality of the data. It mainly focuses on responsibility, ownership, and preventive data quality control.

Data Quality Audits:

Design a plan for conducting quality assurance audits. This involves routine checks of data submitted to the processing centers and management [3].

Technical Data Quality Improvement Solutions

Based on the type of specific data quality issue, the appropriate solution will be selected and implemented here also.

Data Cleansing using ML: It aims to improve the quality of data by identifying and removing errors and inconsistencies [52]. Data cleansing using machine learning (ML) involves leveraging various ML techniques to automate or assist in the process of identifying and correcting errors, inconsistencies, and redundancies in datasets. However, the effectiveness of ML-based data cleansing approaches depends on the quality and representativeness of the training data and the suitability of the chosen algorithms for the specific data cleansing tasks. For instance to handle Missing Value, ML techniques such as regression, decision trees, or k-nearest neighbors can be employed to predict and impute missing values in datasets, whereas, for Outlier Detection, isolation forests, k-means clustering, or density-based techniques can be used to detect outliers in the data. So based on the type of data quality problem, we can choose and implement a suitable algorithms for clean our data.

Using NLP's Fuzzy string matching Algorithm: Fuzzy string matching is a technique used to find approximate matches between two strings [51]. It is used to compare strings that may not be an exact match.

These algorithms employ various techniques to determine the similarity between two strings, taking into account factors such as character substitutions, insertions, deletions, and transpositions. The algorithm would typically involve comparing each string in a dataset with

every other string to identify potential duplicates. However, instead of requiring exact matches, the algorithm would tolerate a certain degree of variation, allowing for differences due to typos, misspellings, or other discrepancies. It uses method of Levenshtein distance measures and the Jaccard similarity coefficient

By setting a threshold for acceptable similarity, you can identify and remove duplicate strings from the dataset. Strings that fall below the threshold are considered duplicates and can be eliminated or merged as needed. This process helps to clean and reduplicate datasets, ensuring that only unique and relevant information is retained.

Data Integration: Data about the same item often exists in multiple databases. This data can take virtually any form (customer name and address data, product data, etc). Data quality management / integration can be applied to virtually any data problem [3]. In the case of Insurance company(AIC), for instance Customer name can be registered at different company's branches and for different Insurance cover(like for Life and Nonlife).

6. Verify

Verification phase is essential for ensuring that, the solutions designed during the Design and development phase effectively address the identified data quality issues and meet the predefined objectives.

In this case, after successful implementation of appropriate solution options for defined specific data quality problem, the effectiveness the model design is will be demonstrated and validated. This stage of the model will be clearly shown in chapter 6 of this paper, with case study approach and evaluated using certain criteria. The designed solution will also be redesigned, based the result after implementation and feedback from relevant stakeholders, if any as stated previously.

Finally, concluded by communicating the overall process work of the solution to relevant stakeholder and getting feedback for future improvement, this means that, it is the final output of proposed model.

Chapter 5: Demonstration and Evaluation

5.1. Overview

A case study approach along with illustrative scenario is used to demonstrate the applicability of the proposed Data Quality Improvement Process model and process validation in accordance with the established solution objectives and requirements. The reason of selecting such kind of approach is that, as explained in chapter one of the paper, the purpose of this research is to improve data quality issues in case of Awash Insurance Company's Information system, through the use of an appropriate systematic data quality Improvement strategy (in this case, Insurance Data Quality Improvement Process model), so applying the solution to primarily defined specific data quality case of the AIC will be assumed to show the appropriateness of the model.

Specific data quality problems encountered in Awash Insurance Companies' information systems, are identified with the help of data or information acquired for the purpose the research work. Based on the type of data quality issue needs to be improved, the appropriate option of solution will be preferred and applied from incorporated improvement solution component of the Proposed Model.

The proposed models brief description is given in the next section.

5.2. Proposed Model description

The Proposed model (named as "Insurance Data Quality Improvement Process model"), provides a comprehensive approach to enhance data quality in the insurance business, specifically designed and developed for Awash Insurance Company, by customizing and incorporating two different existing methodologies, Six Sigma's DMADV process methodology with Total Data Quality Management principles.

The Model was created after detail analysis and customization of existing data quality management and Improvement methodologies and models as shown in section 4.4 of this paper.

What makes the developed model unique from the existing generic data quality management methodology and models is that, it is designed and developed in insurance business process context, considers key stakeholder involvement and clearly defined inputs (identified DQ

problem)and expected outputs(in the means of result and communication)from improvement process activities.

5.3. Implementation Plan

To successfully implement the function of designed Insurance Data Quality Improvement Process model, for specific data quality issue of the selected company (AIC's) case or cases, it need to follow some scientific way, which needs to asses and select the appropriate approach.

Case Study research methodology by Yin (2003) was used as a guideline to set up and execute the case study [29].

Yin's (2003) book "Case Study Research: Design and Methods" outlines a comprehensive approach to conducting case study research. The stages described in the book provide a structured framework for researchers to plan, execute, and analyze case studies effectively.

Details description of the stages for our case using Yin's (2003) case study is shown below.

1) Determine case and Define the Research Questions

This case study will be carried out at the Awash Insurance Company that utilizes an integrated information system.

To determine the type of case to be shown, first it is important to understand the basic business process of the company which is briefly described as follows.

As of other insurance companies, its main goal is to offer insurance services for its customers, which needs to collect personal and industry-specific data from customers that is used for its business Information systems and stored in its integrated system's databases as alphabetic or numerical data.

The finding from document analysis shows that, Awash Insurance company has identified core businesses operational process for its Information system as: *Underwriting, Claims, Reinsurance and Finance*, and other as supportive: *Customer Information, Marketing, System Administration, Workflow Management, Document Management & Image Linking and so on*.

Among the identified core business module, underwriting process is used as a basis for others, because, it is the starting point of service providing task.

Underwriting process has the following steps: *customer information registration, proposal preparation, Premium collection and printing policy* to handover to the customer.

So during these processes the data is collected from different sources and entered into the system by system user. For instance personal data like customer names, residences, phone numbers, medical histories, and accident records is required during customer registration process, and risk related data like, type of risk, sum insured, plate number, chassis number, customer claim history record and similar data will be captured during proposal preparation stage.

The data collected, entered and used during underwriting processes, significantly affects the premium amount to be collected, the claim settlement process, the reinsurance and financial process activities of the company. Hence it impacts the company's decision making process.

So, based on the defined case, it is important to identify the main cause and specific data quality problems of Awash Insurance Company's Information system which helps to answer the first and second research questions of the paper.

2) Design and Selection of the Case or Cases

The case study design describes the sort of case study (i.e. multiple or single case, holistic or embedded) and the case study protocol (the procedure of conducting the case study, including the time frame and planning and the selection of participants) (Van Wierst, J. J. 2019).

In our case, the type of case we are going to handle is, **a single holistic case study** approach aligns well with the objectives of the thesis, allowing for an in-depth exploration of data quality issues within Awash Insurance S.C. and the development of practical recommendations for improvement.

In addition to that, the reason for selecting this kind of approach is that, the study focuses on a specific and single case with comprehensive examination of the case, exploring various dimensions of data quality in Awash Insurance Company's Information system.

3) Preparation of the Data Collection Strategy

During this phase, it is necessary to identify particular areas of data quality problems in the company's (AIC) information system, mostly using the primary data acquired for the study. Secondary data is used as additional input. The researcher gathered primary data from interviews

and document analysis. In addition to this, raw data generated from the Awash Insurance Company's system software, will be used to simulate models.

4) Data Collection

Implication from the finding of the interview from relevant stakeholder shows that, the data collected and entered into the system during underwriting activities mostly affects data quality issues of the company, next to that during claim process the data collected and used for claim processing has also significant effect on every organizational decision making process. In addition to this, from the observation of raw data generated from the company's system software, there is clear visibility of previously mentioned data quality problems.

Following that, the case of Customer Information and risk detail registration process during Company's (AIC) Underwriting Process will be selected as the type of cases used for demonstration purposes, because of its high potential effect on the overall company's business process.

During the customer information registration process, the customer's name may be registered multiple times (with different spellings), the phone number might not be in standard form, and the age of the insured (Assured in the case of Life and Health Insurance) may be incorrectly captured, resulting in instances duplication, inconsistency, inaccuracy, and incompleteness of data quality. As a result, it has an impact on how the company makes business decisions

For Risk to be insured, details like, the type of risk identification, risk details (for instance in motor Insurance cover: sum insured amount, age of the vehicle, Plate and Chassis number, name of driver etc) needs to be perfectly captured, unless and otherwise, the consequence of its incompleteness, Inaccuracy and Inconsistency will affect, the premium to be collected and its claim settlement process, consequently effects the financial business process of the Company (AIC).

5) Analysis of the Data

This step involves systematically organizing, categorizing, and interpreting the collected data to identify patterns, themes, and relationships relevant to the research questions. In recommends using within-case analysis to understand each case individually and cross-case analysis to compare and contrast findings across cases.

In this case, the data collected from primary sources, interview and company's document related to data management were analyzed to drive relevant insight for the model demonstration. In addition to this, relevant raw data with data quality issues generated from company's software were observed and analyzed with stakeholder (Data Analysts) for model simulation.

6) Case Study Report Preparation

This is the final stage, in which the finding and conclusions of the case study (demonstration stage) will be comprehensively communicated and reported.

The report will present and examine the case(s) in detail, as well as address the research questions and objectives. It is well-organized, clear, and supported by data analysis proof.

5.4. Execution of Demonstration for Data Quality Improvement

In this section, it is attempted to show four cases, each the first two cases is technically solvable and the other two cases are that of none technically solvable.

a) Start(Input)

The initial stage in the proposed model is START.

Input = raw data with data quality problem:

In this case *Customer Information* data generated from AIC's system software

b) Define

The stage were Insurance business process, scope of data quality enhancement activities, stakeholder identification and identification of specific data quality issues is clearly defined and specified. As per that, this produces the following outputs:

- Awash Insurance company's conceptual business process is shown using diagram in Appendix IV.
- Scope of the Insurance data quality Improvement Process activities

As mention in 5.3(1) section, the improvement process will be applied on one of the core business process of AIC which is *underwriting*, because initial data entry process for the rest of core and supportive business process of the company begins at this stage and mostly affects the company's decision making process.

- Identification of relevant stakeholders

As mentioned in section 4.1.4 of this paper, three types of stakeholders namely,

- Data Analyst, System user (specifically Underwriter) and the researcher itself has been selected as the relevant stakeholder in the data quality improvement process.

- Identification of specific data quality issues.

Now we have identified four types' data quality problems in customer Information data and Risk detail data as the following:

- Customer name duplication named as- Instance duplication issue
- Phone number of customers are not in standard format –Inconsistency issue
- Missing data- Incompleteness issue
- Incorrect data – Accuracy Issue

The following image below shows customer Information data named as *CIF Corporate Report* with DQ problems, which is primarily used for our model simulation purposes.

Table 2: Customer Information data with Data Quality Issue

		Awash Insurance S.C.				User Name ***	
		CIF Corporate Report				Date	09/04/2024
						Time	09:26:23 AM
CIF Number	Customer Name	Branch Name	Custom	Address	Phone	Mobile	E-Mail ID
CIFFFN**2506423	Client 1	Finfinee Grand Main Branch	General	200,,Addis		911949***	
CIFFFN**32519723	Claint 2	Finfinee Grand Main Branch	General	New,,Addis	0911-2284**	911401***	
CIFFFN**62533023	Client 2	Finfinee Grand Main Branch	General	New,,Addis	0911-2284**	911949***	
CIFFFN**92546323	Claint 3	Finfinee Grand Main Branch	General	NA,,Addis		911401***	
CIFFFN**12255962	Client 3	Finfinee Grand Main Branch	General	new,,Addis		911949***	
CIFFFN**15257292	Claint 4	Finfinee Grand Main Branch	General	298,,Addis		911401***	
CIFFFN**18258622	Client 4	Finfinee Grand Main Branch	General	new,,Addis		911949***	
CIFFFN**21259952	Claint 5	Finfinee Grand Main Branch	General	new,,Addis		911401***	
CIFFFN**24261282	Client 5	Finfinee Grand Main Branch	Major	New,,Addis		911949***	
CIFFFN**27262612	Claint 6	Finfinee Grand Main Branch	General	new,,Addis		911401***	
CIFFFN**30263942	Client 6	Finfinee Grand Main Branch	General	143,,Addis	09770820**	911949***	
CIFFFN**33265272	Claint 7	Finfinee Grand Main Branch	General	New,,Addis		911401***	
CIFFFN**36266602	Client 7	Finfinee Grand Main Branch	General	New,,Addis	2511166748**	911949***	
CIFFFN**39267932	Claint 8	Finfinee Grand Main Branch	General	NEW,,Addis		911401***	
CIFFFN**42269262	Client 8	Finfinee Grand Main Branch	General	364/367/3F,,Addis		911949***	
CIFFFN**45270592	Claint 9	Finfinee Grand Main Branch	General	New,,Addis		911401***	
CIFFFN**48271922	Client 9	Finfinee Grand Main Branch	General	400/2/1,,Addis		911949***	
CIFFFN**51273252	Claint 10	Finfinee Grand Main Branch	General	new,,Addis		911401***	
CIFFFN**54274582	Client 10	Finfinee Grand Main Branch	General	new,,Addis		911949***	

c) Measure

The third step of the proposed Insurance Data Quality Improvement Process Model's is **Measuring** activity.

In this case, for selected data set, we are going to measure the state of data quality in terms of its predefined data quality dimension:

So here we have identified

- How many duplicate instance of customer name(with different spelling exists),
- How many mobile phone number of customers are not in standard format(Inconsistency issue), how many missing values are there in risk registration data(Incompleteness Issue) and
- How many values of instance are inaccurate in terms of data context (Inaccuracy Issue)?

d) Analyze

In this stage, the main cause of the defined and measured data quality issues of Awash Insurance Company's Information system will be analyzed as follows.

In addition to the finding of interview,

- Observation and analysis activities of raw data with two relevant stakeholders one is data Analyst and the other is System user were conducted.
- Finally the major cause of this problem is identified and categorized into two major as:
 - internal cause:
 - Caused by user error in customer services (During underwriting process).
 - Lack of Organizational DQ management standard.
 - external Cause:
 - Due to data coming from outside the organization(E.g.: Banks)
 - Due to inconsistent data received from corporate customers.

The activities performed in this phase helps to address the research question number 2, of the paper.

e) Design and Improvement

In this stage, we are going to apply solution options designed and incorporated as a component of this phase of our proposed model (Insurance Data Quality Improvement Process), based on

the specific type of data quality issues identified in the Define phase of this model, we have identified four illustrative scenarios.

As mentioned in section 4.4.2 of this paper:

The corresponding solution option from this, phase will be implemented to specific Data Quality problem as follows:

- i. **Scenarios 1:** Handling or improving of **Instance duplication** issue (Decided to be *technically solvable*).

Here in this case, as shown in above corporate customer data there is duplicate instance of data. For example customer name 'Ethio Legal Shiled ' is encrypted as Clint1 and Cliant2 is for 'Ethio Legal Shiled LLP', So the customer name is the same but registered twice, so it is duplicated instance with different name,

As shown in Figure 5 of Insurance data quality Improvement Process model, in design and Improvement phase, there is an option called **Fuzzy string matching**, which is appropriate for these specific data quality problems (Instance duplication).

As mentioned in section 4.4.2, applying **Fuzzy string matching** solution removes the duplication issue and hence fixes the problem and its implementation logic is described as follow:

Fuzzy string matching, in which the degree of its similarity is determined with fuzzy ratio, so if the fuzzy ratio is more than 70 %(threshold default value), then their corresponding customer's phone number of is checked. If they have the some phone number, the probability of determining that it is one customer is extremely high, following this, those highly similar string as are merged to a single name, hence duplication issue will be solved.

The python script below is shows...

```
python

from fuzzywuzzy import fuzz
from collections import defaultdict

def find_duplicates(customer_names, threshold=70):
    .. ..
```

Figure 6: python script code for Instance duplication removal

- ii. **Scenarios 2:** Handling or improving of **Inconsistency** issue (Decided to be *technically solvable*).

Here in this case, as shown in above corporate customer data, phone number of some customers are registered with adding country code and other are not, For example: customer name ‘Population Services International Eth’ has registered with phone number ‘+25111667****’ which is in its standard format, whereas for other customers phone number starts with ‘09****’ which is not in its standard form.

Now in our design and Improvement phase of our Insurance data quality Improvement Process model, there is an option called ‘**Data Cleansing using ML**’ so for this specific data quality problems (Inconsistency issue), applying the solution provides standard format for customer’s phone number.

- iii. **Scenarios 3:** Missing data- **Incompleteness** issue. (Decided to be *Non-technically solvable*).

For this case of data quality issue, we can use risk detail data which contain risk related such type of risk, sum insured amount, claim history record specific place of the risk, etc.

So to handle such data quality problems, the following solution will improve the incompleteness of relevant data for the respective company’s business process.

Training and Education:

In order to prevent human error in data input, customer service representatives and other individuals involved in data gathering must receive training, which means educating staff on data quality best practices and the importance of accurate data entry.

Data Collection Improvement:

Improve data collection processes to minimize the occurrence of missing values in future datasets.

Data Stewardship:

It mainly focuses on responsibility, ownership, and preventive data quality control.

Data Quality Audits:

Design a plan for conducting quality assurance audits. This involves routine checks of data submitted to the processing centers and management [3].

iv. **Scenarios 4:Incorrect data** – Accuracy Issue(Decided to be *Non-technically* solvable)

For this kind of Insurance data quality problem, it recommended to Adopt and customize Panwar, M. (2015)'s .stated as below

“Standards and Protocols.-When data quality problems are encountered importing data into the data warehouse, there are four viable actions that can be taken. We can:

*Reject the error, Accept the error, Correct the error and Apply a default value for the erroneous data “.*we can refer detail of the concept from section 4.4.2.

f) Verify

In this stage, we are going to proof the validity of the improvement process.

In this case, it is shown that the raw customer information data and risk detail data having data quality problems are taken as input for Model simulations. Passing through define, measure and improvement phases of the proposed model (Insurance Data Quality Improvement Process model), data quality issues such as: customer name duplication, Customer phone no(without country code), missing and incorrect values are corrected and improved.

This implies that the model's effectiveness in terms of improving the data quality of AIC. Additionally, in this step, feedback from respective relevant stakeholder can make to redesign the solution model for further improvement.

g) Stop(Output)

The final output from this Data Quality Improvement process as stated as follows:

Output:

- Awash Insurance Company's corporate customer's data free of customer name duplication with their corresponding standardized phone number (add with country code. e.g. +251).
- Risk detail data's missing and inaccurate values will also probably improve with the take measure from the model's improvement solution design.

5.5. Conclusion of the Demonstration process

This section summarizes the key findings, outcomes, and implications of the developed model demonstration activity.

As mentioned in section 5.3, Case Study research methodology by Yin (2003) was used for the execution of the case study that helps to provide a structured framework for researchers to plan, execute, and analyze case studies effectively.

Based on yin (2003) approach, we identified underwriting operation, among core business process, of Awash Insurance as case type, which includes customer information registration activity as initial main task, and which has significant effect on other operational process of the company(AIC),then data collection process(in this case Raw data for model simulation),analysis and identification of specific data quality problems were identified, finally implementation of data quality improvement process by execution of the model was done.

The demonstration phase for the developed model (Insurance data quality Improvement Process model) brought about encouraging results in terms of enhancing data quality within Awash Insurance S.C.'s insurance information system. Through intensive evaluation and approval, it was noticed that the model efficiently identified and improved data quality issues such as instance duplication, inconsistencies, incompleteness and Inaccuracy, as shown through case study approach.

Furthermore, input from stakeholders and end users during the demonstration phase were highly beneficial, demonstrating a high level of acceptance and usability for the established model. This highlights its practical relevance and application in real-world insurance scenarios.

The model's performance under controlled conditions is good, but its scalability and robustness for handling enormous data quantities require further analysis and continuing maintenance to ensure long-term functionality.

In general, the demonstration process provided useful insights into the possibilities for using advanced data quality techniques to improve the performance of insurance information systems.

5.6. Evaluation process

In this chapter we evaluate the way of providing solution to predefined problem in the research, particularly the effectiveness of the developed artifact (In this case, Insurance Data Quality improvement process Model).

As stated in section 1.3 of the paper, the main aim of the research was to improve data quality for Insurance Information system, particularly for Awash Insurance Company. As such the research question is formulated and the way those questions are going to be answered is specified.

The research used Peffers et al.'s (2007) design science methodology for information systems research, which provided clear guidelines for problem, solution development, and evaluation.

As per the provided guidelines, the study identifies a data quality issue of the company's (AIC) insurance information system, reviews its objectives, and designs a model called the 'Insurance Data Quality Insurance Information System'.

The model is demonstrated through a case study with illustrative scenario and the process is evaluated using a qualitative assessment approach since this paper's designed conceptual model is primarily aimed to reflect qualitative characteristics.

Evaluation Criteria's used for evaluation process.

1) Alignment with thesis objectives and research question

As the main objective of the thesis research is to enhance data quality of Insurance information system, it needs to acquire relevant information for appropriate solution to be developed.

In this case, secondary data (from literature) and primary data collection was conducted through interview with relevant stakeholder and document analysis was performed. The finding of these activities helps to addresses *research question number 1 and 2*.

The specific data quality issues existed from reviewed document analysis and raw data shows, there is visible instance duplication, Inconsistency, Incompleteness and Inaccuracy issues of data quality issues, which answered the *first research question*.

Once the cause and specific data quality issues are identified, appropriate Improvement solution in the form of model is designed and developed and then, proofed with demonstration activity which brought about encouraging results in terms of enhancing data quality within Awash Insurance S.C.'s insurance information system; this addresses *research question number 3*.

So following from the above finding, the proposed solution aligns with thesis objectives and research question of the study.

2) Practicality of Solutions

Once the specific data quality issues of Awash Insurance S.C.'s insurance information system is identified, it is shown that Model implementation on real world data (Raw data generated from AIC's system software) was demonstrated with case study activity. For instance, Customer name duplication issue is successfully improved through applying technical DQ improvement solution (Fuzzy string matching of NLP application) and other data quality issue are also improved as shown is demonstration section of the paper. In addition to this, discussion with stakeholders (Data Analysts expert and researcher itself) regarding practicality issue is performed.

The finding from detail discussion with stakeholders implies, the model implementation is easy and practical, however to ensure its comprehensiveness, the scope of its implantation has to be extended to the company's (AIC) other core, as well as supportive business processes.

3) Ethical considerations

The researcher obtained permission from Awash Insurance Company and Addis Ababa University School of IS to collect data, ensuring data privacy and confidentiality. The study was informed to participants and Personal information de-identification approach to protect their identities, which shows the thesis work, is highly responsible research.

4) Scalability and Sustainability

The proposed model for the DQ issue of AIC's can easily be implemented and extended to similar data quality issues of other Insurance companies as well, however it needs continuous monitoring and improvement for better performance.

Generally finding of evaluation process using the above evaluation approach, implies the appropriateness of the proposed model and encouraging results in terms of enhancing data quality

Chapter 6: Conclusion and Recommendation

6.1. Conclusion

This thesis has addressed the critical issue of data quality within the context of insurance information systems, specifically focusing on Awash Insurance S.C, through the development and application of the Insurance Data Quality Improvement Process Model.

The proposed model is created as an artifact for Awash Insurance Company's Information system's data quality problem, after detail analysis of existing Data Quality management methodologies through customizing and incorporating two different existing methodologies, Six Sigma's DMADV process methodology with Total Data Quality Management principles.

This Model (Insurance Data Quality Improvement Process Model) provides a comprehensive framework for assessing and enhancing data quality, targeting common issues such as duplication, inconsistency, incompleteness, and inaccuracy.

Our findings demonstrate the significance of proactive data quality management in ensuring the reliability and effectiveness of insurance operations. By implementing the proposed model, organizations like Awash Insurance S.C. can identify and rectify data quality issues, thereby improving decision-making processes, customer satisfaction, and overall business performance.

Moreover, the case study conducted at Awash Insurance S.C. has illustrated the practical applicability and effectiveness of the proposed model in a real-world setting. The successful implementation of data quality enhancement strategies highlights the potential benefits of focusing on data governance and quality assurance.

6.2. Recommendation

It is important to acknowledge the limitations of this research, including the specific context of Awash Insurance S.C. and the generalness of the proposed model to other insurance organizations. Future research should explore validating the model across various insurance sectors and enhancing data quality management dimensions like timeliness and relevance.

In general, the proposed Insurance Data Quality Improvement Process Model offers a valuable framework for addressing data quality challenges within insurance information systems, with implications for enhancing operational efficiency, regulatory compliance, and strategic decision-

making. One of the interesting points regarding the proposed model is that, it is easily customizable, extendable and improvable. Moreover, Insurance organizations can achieve sustained success in an increasingly data-driven industry by adopting a proactive approach to data quality management.

References

1. Sidi, F., Panahy, P. H. S., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012, March). Data quality: A survey of data quality dimensions. In 2012 International Conference on Information Retrieval & Knowledge Management (pp. 300-304). IEEE.
2. Blomqvist, N. (2019). Data quality methodologies and improvement in a data warehousing environment with financial data.
3. Panwar, M. (2015). Application of six sigma for data quality improvement in an insurance company. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, 4(09), 63-68.
4. Gorla, N., Somers, T. M., & Wong, B. (2010). Organizational impact of system quality, information quality, and service quality. *The journal of strategic information systems*, 19(3), 207-228.
5. Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*. Springer.
6. A. Mitra, "Classifying Data for Successfull Modelling," 2012. [Online]. Available: www.dwbiconcepts.com.
7. C. Batini and M. Scannapieco, *Data Quality - Concepts, Methodologies and Techniques*, Springer, 2006.
8. R. D. Reid and N. R. Sanders, *Operations Management*, 4th Edition, Wiley, 2009.
9. "Business Dictionary," WebFinance, Inc., [Online]. Available: <http://www.businessdictionary.com/definition/quality.html>.
10. J. M. Juran and A. B. Godfrey, *Juran's Quality Handbook*, 5th Edition, New York: McGraw-Hill, 1999.
11. M. Chen, D. Ebert, H. Hagen, R. S. Laramée, R. van Liere, K. L. Ma, W. Ribarsky, G. Scheuermann and D. Silver, "Data, Information, and Knowledge in Visualization," *Computer Graphics and*
12. R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, 1996.

- 13 ELFAKHFAKH, M. T. M. O. (2019). Big data quality: stock market data sources assessment.
- 14 D. Loshin: “The Practitioner’s Guide to DQ Improvement”, Elsevier, published by Morgan Kaufmann Publishers, October 2010.
- 15 T. Redman, Data Quality for the Information Age, 1997.
- 16 R. Y. Wang, V. C. Storey and C. P. Firth, "A Framework for Analysis of Data Quality Research," *IEEE Transactions on Knowledge and Data Engineering*, 1995.
- 17 R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, 1996.
[27]
- 18 R. Y. Wang, "A Product Perspective on Total Data Quality Management," *Communications on the ACM*, 1998.
- 19 Y. W. Lee, D. M. Strong, B. K. Kahn and R. Y. Wang, "AIMQ: A Methodology for Information Quality Assessment," *Elsevier, Information & Management*, 2002.
- 20 R. Y. Wang: “A product perspective on total DQ management”, in Communications of the ACM, vol. 41, no. 2, pp. 58–65, Feb. 1998.
- 21 C. Batini, D. Barone, M. Mastrella, A. Maurino and C. Ruffini, "A framework and a methodology for data quality assessment and monitoring," in *12th International Conference on Information Quality*, 2007.
- 22 Y. Wand and R. Y. Wang, "Anchoring Data Quality Dimensions in Ontological Foundations," *Communications of the ACM*, 1996.
- 23 J. S. Ryu, J. S. Park, and J. H. Park: “A DQ Management Maturity Model”, in ETRI Journal, vol. 28, no. 2, pp. 191–204, April 2006.
- 24 C. Batini and M. Scannapieco, Data Quality - Concepts, Methodologies and Techniques, Springer, 2006.
- 25 ELFAKHFAKH, M. T. M. O. (2019). Big data quality: stock market data sources assessment.
- 26 C. Batini, C. Cappiello, C. Francalanci, A. Maurino: “Methodologies for DQ assessment and improvement”, in ACM Computing Surveys, vol. 41, no. 3, pp. 1–52, July 2009.
- 27 .VomBrocke, J., Hevner, A., &Maedche, A. (2020). Introduction to design science research. Design science research. Cases, 1-13.
- 28 Mishra, S. B., &Alok, S. (2011). Handbook of research methodology.

- 29 Van Wierst, J. J. A Process Model for organizational Data Quality Assessment.
- 30 vanWierst, J. J. A Process Model for organizational Data Quality Assessment.
- 31 Nikiforova, A. (2020). Definition and Evaluation of Data Quality: User-Oriented Data
Object-Driven Approach to Data Quality Assessment. *Baltic Journal of Modern
Computing*, 8(3).
- 32 FikiruNegasa (2015), "*Health management information system data quality and
information usage improvement project: case of begi district*".
- 33 ShewayeHailu (2017), "*Improving data quality and information use in maternal and child
health departments of health centers: the case of Dessie city administration, south wollo,
amhara region*"
- 34 RegasaBayisa (2014), "*Assessment of health management information system (HMIS) data
quality and information use: The case of Yekatit 12 Hospital, Addis Ababa*".
- 35 R. Mahanti, "Data Governance Implementation: Critical Success Factors," *Software
Quality Professional*, vol. 20, no. 4, 2018.
- 36 S. Robertson and J. Robertson, *Mastering the Requirements Process*,
3rd ed. Addison-Wesley, 2014, isbn: 978-0321815743.
- 37 Nikiforova, A. (2020). Definition and Evaluation of Data Quality: User-Oriented Data
Object-Driven Approach to Data Quality Assessment. *Baltic Journal of Modern
Computing*, 8(3).
- 39 Fang, X., & Anderson, J. (2024). *Importance of Master Data Management (MDM) in
Enterprise Data Governance* (No. 12472). EasyChair.
- 40 Wendelborn, C., Anger, M., & Schickhardt, C. (2023). What is data stewardship? Towards
a comprehensive understanding. *Journal of Biomedical Informatics*, 140, 104337.
- 41 Buitenhuis, V. (2023). Designing a holistic method for enhancing data quality with the use
of machine learning: A master thesis for ICT in Business & the Public Sector at Leiden

University.

- 42 SOYER, A. (2013). Data Quality Management: A Solvency II Perspective.
- 43 Kabir, J. M. (2016). *Factors influencing customer satisfaction at a fast food hamburger chain: the relationship between customer satisfaction and customer loyalty*. Wilmington University (Delaware).
- 44 Akheel, M. (2020). Improving Data Quality for the Insurance Industry Why and How.
- 45 Freudiger, J., Rane, S., Brito, A. E., & Uzun, E. (2014, November). Privacy preserving data quality assessment for high-fidelity data sharing. In *Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security* (pp. 21-29).
- 46 Castillo, I., Losavio, F., Matteo, A., & Bøegh, J. (2010). REquirements, Aspects and Software Quality: the REASQ model. *J. Object Technol.*, 9(4), 69-91.
- 47 Gualo, F., Caballero, I., Rodríguez, M., & Piattini, M. (2023). A Data Quality Model for Master Data Repositories. *Informatica*, 34(4), 795-824.
- 48 Aljumaili, M. (2016). *Data quality assessment: Applied in maintenance* (Doctoral dissertation, Luleå tekniska universitet).
- 49 Otto, B., Wende, K., Schmidt, A., & Osl, P. (2007). Towards a framework for corporate data quality management.
- 50 Lizot, E., & Islam, S. M. (2018). The impact of Privacy concerns in the context of Big Data: A cross-cultural quantitative study of France and Bangladesh.
- 51 Pikies, M., & Ali, J. (2021). Analysis and safety engineering of fuzzy string matching algorithms. *ISA transactions*, 113, 1-8.
- 52 Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4), 3-13.
- 53 Abedjan, Z., Golab, L., & Naumann, F. (2017, May). Data profiling: A tutorial. In *Proceedings of the 2017 ACM International Conference on Management of Data* (pp.

1747-1751).

- 54 Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, 41(3), 1-52.

Appendix

Appendix I: Interview question Checklist

List of Interview Questions

Pre-Interview protocol, for example: *Thank you for agreeing to participate in this interview.*

*I'm conducting research on **Improving Data Quality for Awash Insurance's Insurance Information System.***

- 1) Are you familiar with the concept of data quality and its role in business operations and decision-making processes of Insurance organizations, specifically at Awash Insurance S.C.?
- 2) What are the current data quality issues or challenges faced by Awash Insurance S.C. in its information system?
- 3) What do you think about the potential causes of data quality problems in Insurance sector, specifically the Awash Insurance Company's information system?
- 4) How do these data quality issues impact the overall efficiency and effectiveness of the insurance processes within the organization?
- 5) What role do stakeholders, such as employees, customers, and external partners, play in ensuring data quality within Awash Insurance S.C.?
- 6) What resources or support does Awash Insurance S.C. require to effectively address data quality issues within its information system?
- 7) In your opinion, what are the key priorities or areas of focus for improving data quality within the insurance information system at Awash Insurance S.C.?
- 8) If the solution will going to be offered for data quality problems within Awash Insurance S.C, how its effectiveness will be evaluated?
- 9) Do you have any additional comment or remark?

At the end of the interview: *"Thank you for taking the time to conduct this interview!!"*

Appendix II: Checklist for Organizational document analysis.

The following lists are, the type of documents planned to be reviewed and analyzed for Additional input.

- 1) **Policy Documents:** such as Data management policies and procedures, Quality assurance policies, Compliance regulations related to data quality in the insurance industry, if any
- 2) **Data Quality Reports:** data quality assessments or audits, Identified data quality issues and their impact on business operations of Awash Insurance S.C.
- 3) **System Documentation:** Current information system architecture of the company (AIC), training manuals or guides for employees.
- 4) **Data Improvement Initiatives:** Project plans, progress reports, and outcomes, if any
- 5) **Risk Management Documentation:** Risk assessments related to data quality vulnerabilities, if any.

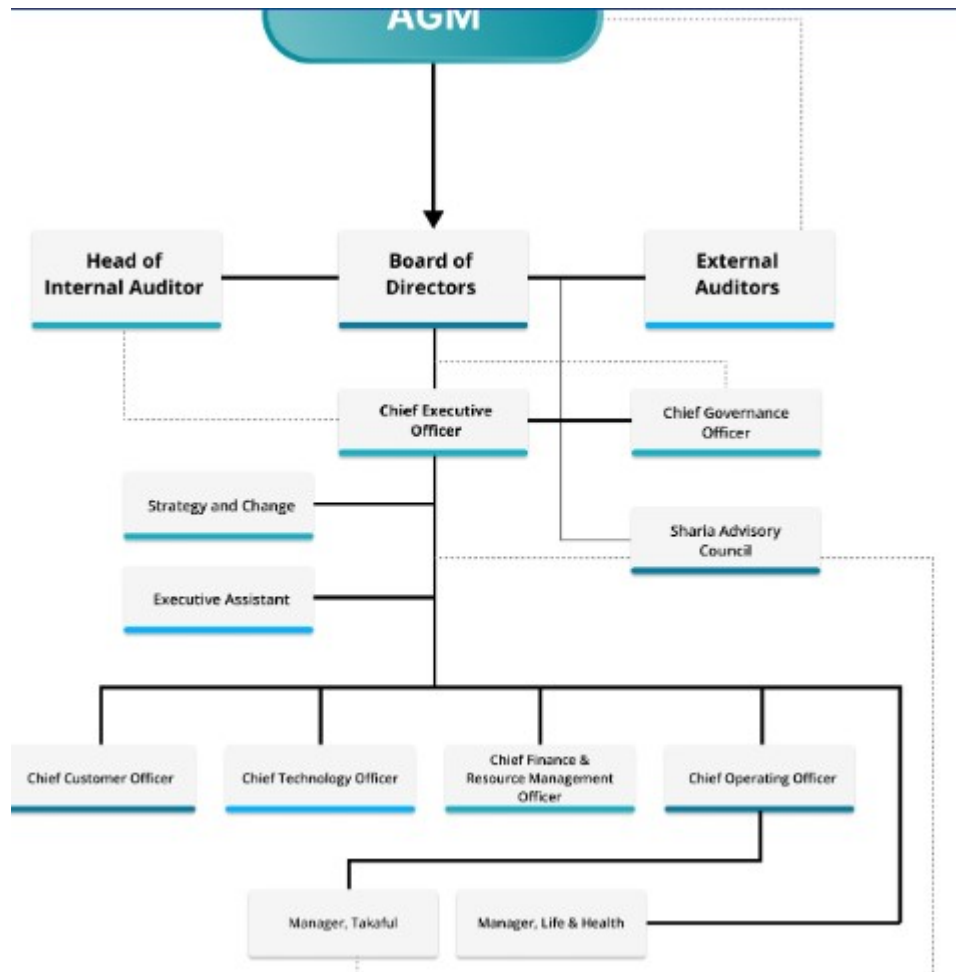
Note: *I think it needs to be reviewed and refined based on the feedback from the relevant stakeholders.*

Appendix III: Search terms used for literature review

Activities
Approach
Assessment
Business
Company
Data
Dimensions
Framework
Guide
Information
Measurement
Measures
Method
Methodology
Metrics
Organization
Organizational
Process
Quality
Roles
Tasks

Appendix IV: Awash Insurance Organizational structure

(Used to understand the company's business process)



Appendix V: One of System Manual used for Document Analysis

1. General Insurance Information System(GIIS) Overview

- GIIS is role based General Insurance Software – Users only access GIIS features which are related to their work.
- GIIS has 9(nine) modules - four are core modules and five are supportive modules
- **Core modules** – Underwriting, Claims, Reinsurance and Finance module
- **Supportive modules** - Customer Information, Marketing ,System Administration, Workflow Management, Document Management & Image Linking

2. Customer Information File(CIF)

There are two types of customers – Individual & Corporate

2.1. Create Individual Customer –Customer Information→Entries→CIF Individual

2.2.Create Corporate Customer–Customer Information→Entries→CIF Corporate

3. Underwriting Module

Our Goal: To Print Policy and handover to a Customer

STEPS:

1. To Print Policy and Handover to Customer:

1.1.**Prepare Proposal:** Underwriting→Transactions →Proposal→Select Policy

Name & enter Policy details→Save

- Upon Save- Proposal Number is generated
- Computation Sheet can be viewed upon Save in Step 1.1. OR it can be viewed in Inquire mode after the Proposal is approved.
- **Scenario** : Policy Name=....., Risk Name=..., SI=

Important: Business source selection– don't forget to select the business source

1.2.**Authorize the Proposal** : [This will be the role of Branch Managers]

- Login In as Branch Manager
- Click on Authorization Basket
- Approve/Return For Edit/Reject with Remark for correction

1.3.**Generate Policy Number:** Underwriting→Transactions →Print Policy→Enter Proposal Number→Click on Search→Click on Generate Policy Number

- This will generate policy Number
- This step should also be followed during Policy Renewal. Note that during policy renewal new policy number will not be generated.

1.4. **Premium Receipt Order** [This is Underwriters Role]

Underwriting → Transactions → Premium Receipt Order → Select Premium From → Select the Policy Number for which the Premium is collected → Complete the Accounting Details and Save.

Important: Ask the Customer mode of payment and WHT information before you prepare premium receipt

1.5. **Collect Premium** (This is Cashiers Role)

1.5.1. Mode of Receipt Can be in Cash or Cheque

1.5.2. If mode of Receipt is in Cash, go to General Ledger → Cashier Transaction → Open Cash and Collect Premium (Premium Receipt).

1.5.3. If mode of Receipt is in Cheque, and then go to General Ledger → Transactions → Outward Collection / Inward Remittance → Select the Bank Name in which you **deposit the Cheque** and click on Go.

1.5.4. Print receipt, sign on it, put stamp and handover to customer – for both mode of receipts.

1.5.5. After some time Clear the Cheque : General Ledger → Transactions → Cheque Return

Note: Note copy of ***Premium Receipt*** is available at General Ledger → Transaction Reports-Grouped → Premium Receipt Voucher

Caution: Don't save premium receipt order before you collect the premium amount (be it is in cash or cheque).

1.6. **Print Policy and Handover to Customer** (This is Underwriters Role):

Underwriting → Transactions → Print Policy → Inquire → Entry Policy Number → Click on Print Policy.

Important: Don't skip printing policy. This step is very important and has a lot of implication related to Agent/Broker commission computation, Reinsurance allocation, Policy Renewal etc.