Intelligent Techniques for Data- Information- Knowledge Evolution

THESIS

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Artika Agrawal

Graduate Program in Computer Science and Engineering

The Ohio State University

2011

Master's Examination Committee

Jay Ramanathan, Thesis Advisor

Rajiv Ramnath, Academic Advisor

Abstract

The quality of data in an Enterprise Data Warehouse (EDW) plays a very important role in decision making activities like cross selling products and services or to identify unmet market needs etc. Every organization puts in a lot of resources to maintain data of high quality, in order to make informed decisions regarding market trends or to create business strategies. A lot of work is done to identify the types of data quality issues occurring in data warehouses and a large number of industry best practices and standards have been laid down to improve data quality. However, not much work is done for continuous data quality management and improvement particularly in customer data domain.

This work begins with the identification of 1) various types of data quality errors in a customer-centric financial database that are introduced in an on-going bases 2) identifies the root causes of those error types and 3) the corrective actions. We have also proposed a framework for continuous improvement and governance of EDW that achieves greater levels of traceability and decision making across four different organizational levels- Infrastructure, Operations, Business and Strategy. The database used for this purpose is a real-world EDW which is being fed by multiple legacy source systems that continuously introduce redundant customer instances causing data duplication. To address this problem of duplicated data we have used neural networks for customer classification which ensure

continuous data quality improvement. We also explore more generic applications of data quality on knowledge extraction.

Dedication

I dedicate this work to my friends, family and my dogs- Moti, Buzo, Nancy, Christy and Champ.

Acknowledgements

A special thanks to my advisors Jay Ramanathan and Rajiv Ramnath for their continuous efforts and guidance in this work. I also want to thank Tara Paider and Tom Paider for their continued support and mentoring.

I thank my friends at CETI and Nationwide Insurance for providing me the data and other information related to business processes and policies.

Vita

2004…….....Central Board of Secondary Education, India

2005-2009...……....B.Tech., CSE, Uttar Pradesh Technical University, India

Fields of Study

Major: Computer Science and Engineering

Table of Contents

Abstract ii

Dedication iv

Acknowledgements v

Vita vi

Table of Contents vii

List of Figures x

List of Tables xi

Chapter 1: Introduction 1

1.1 The On- Going Data Quality problem 1

1.2 Effects of Data quality 3

1.3 Challenges 5

1.4 Thesis Outline 5

Chapter 2: Terminology 7

Chapter 3: Related Work 11

3.1 Enterprise Architecture 11

3.2 Lean Thinking 13

3.4 Data Mining 16

3.5 Balanced Scorecard: 17

Chapter 5: Case Study- Nationwide Insurance 21

5.1 Planning to Achieve High Data Quality 24
5.2 The Process of On- Going Quality Problem 26
5.3 Corrective Actions 27
5.4 Neural Network based Customer Classification 28
Chapter 6: Research and Contributions 32
6.1 Problem analysis, classification and quantization 32
6.3.1 Data Collection 32
6.1.2 Existing Error Types 34
6.2 Root cause analysis 36
6.3 Data Mining- Usefulness and Limits 38
6.4 Recommendations for Data quality Improvement 42
6.5 Classification of customers using Machine Learning 45
6.5.1 Splitting Data for the Experiment 47
6.5.2 Feature Vector Extraction 48
6.5.2.1 Binary Distance 48
6.5.2.2 Euclidean Distance 48
6.5.2.3 Levenshtein Distance 49
6.5.3 Experimental setup: 49
6.5.4 Results and Conclusion for Customer Classification 50
Chapter 7: Governance related to Information and Knowledge 54

Chapter 8: Effectiveness 61

Chapter 9: Future Work 62

References 63

List of Figures

Figure 1. The Data- Information- Knowledge ladder 4

Figure 2. A Balanced Scorecard..... 20

Figure 3. Interactions between the central repository and various data touch points 23

Figure 4. A Neural Network [22] with one hidden layer and 20 neurons..... 29

Figure 5. A Neural Network [22] with two hidden layer and 20 neurons per layer 29

Figure 6. *A Sigmoid function..... 31

Figure 7. Starlight Generated Figure for error classification based on customer attributes and source system generating the data 40

Figure 8. Starlight Generated Figure for error classification based on stakeholder generating the data 41

Figure 9. An Enterprise Framework for Process Governance 56

List of Tables

Table 1. List of existing error types… 34

Table 2. Results for Binary distance dataset 51

Table 3. Results for Euclidean distance dataset 52

Table 4. Results for Levenshtein distance dataset..... 52

Chapter 1: Introduction

Before we explain our approach to solve the data quality problem we should first understand what are the causes of the data quality problems and what effects does bad data quality have on the proper functioning of an organization.

1.1 The On- Going Data Quality problem

Data quality can be described as the absence of any or all of various characteristics of data like completeness, accuracy, timeliness, exclusiveness etc in the data. When a number of legacy database systems are acquired by organizations, the quality of data is at a huge risk. This is because different standards and representations followed and used by different database systems can cause data quality issues. Data is collected from various legacy database systems and is fed to a central repository. There are multiple touch points where this data in the central repository is manipulated like front end systems used by agents, Customer service representatives, data analytics etc. Unfortunately, any changes made to the data in this central repository are not migrated back to the source systems. So when a weekly or nightly batch process takes place the data of the central repository is again changed back to a new value or an old value of the legacy database system and the previous changes made to the data in the central repository are lost. The previous changes

made to the central system might have been the most up-to-date changes for customer data, which are now lost due to the batch updates from the source system. In the worst case, this particular problem can reoccur with each batch update.

The data warehouse implementation in most of today's real world applications do not have a common data staging environment where we can view the data before being loaded into the target database system but only after the required transformations have been done on it, so that we can cross check the transformations on the incoming data and if needed, revert back the changes before it is committed to the target database system. The data quality problem becomes worse due to the inconsistency across different legacy database systems and also because the data updates made to the central repository at various touch points are not enforced back to the source systems. Also, the parsing rules to classify customers based on the different attributes are inefficient and do not evolve over time with different databases getting added to the central system with businesses merging together. This results in multiple instances for the same customer being created in the central system. This can cause delayed and bad services for the customer and wrong analytics for business such as wrong prospective customers to cross sell other products and services. For instance consider a scenario where a customer calls the customer service to inquire about the discount he will get on product C given that he already owns two other products A and B from the same organization. When the Customer service representative searches for the customer in the central repository the result retrieves only one instance of the customer where he just owns product A and not the instance where he owns product B. This is because the customer information for

product A is under the name Alice Adam while the customer information linked with product B is A Adams. This difference in the customer information can be due to various reasons. For e.g. this can be because the Extract Transform and Load (ETL) process which takes the data from source systems and feeds it to the central repository does a truncation on the first name of customer for product B or because the field length for source system for product B is small enough to not store the complete first name of the customer. This will cause both customer disappointment and loss of business. Whereas if the customer was provided better and timely service with accurate customer information it would have earned customer interest along with more products and services sold. This is just one example of what bad effects low data quality can have on your business. On the other hand better customer service can spread the word and improve the reputation of the organization resulting in customer retention and increased customer base.

1.2 Effects of Data quality

"Data on its own does not speak" but once it is mined for the needed information, it is a huge source of knowledge. What if the data is incorrect, incomplete, missing or garbled? Information extraction from such data will only cause loss of time and resources resulting in uninformed and incorrect decisions at organizational level to develop strategies and to foretell the future and market trends. High quality data is very important for every organization today to draw inference for developing business and marketing strategies. All over the USA businesses lose more than $600 billion yearly according to one of the reports from TDWI (The Data Warehousing Institute).

Figure 1. The Data- Information- Knowledge ladder

Data: Data is just a description about the world around us. It is the raw facts about the world.

Information: When this data is organized and refined it becomes information. It has now evolved and matured and is useful.

Knowledge: Knowledge is getting a better understanding of these facts and information. Information becomes knowledge when actors apply intelligence and insights to the information. “Knowledge is defined as the meaningful links with which people make in their minds between information and its application in action in a specific setting” or “Knowledge also is information with meaning”.

For e.g. When a customer enters some values into a form, when he buys some product or service, it is factual *data* about the customer. However, when a customer calls the customer service representative for an enquiry, the customer data provides *information* about the customer to enable the customer service representative. But when we use some data mining tools or human intelligence on this information we can get *knowledge* to predict customer preferences or market trends.

1.3 Challenges: The data quality problem is not a one time problem. It can grow worse with new legacy systems coming in and feeding the central repository. Even if the existing problems are fixed, new kind of errors can arise over time. We need methods to track data quality and evaluate it in an intermittent fashion. The data is not consistent across systems and is not well labeled so as to infer the data quality issues within it. For e.g. there is no way of telling if a customer instance does not match the other customer instance due to the missing address etc. Some challenges in this work are:

- Inconsistency of data across systems
- Unlabeled and ambiguous data
- What continuous data quality improvement methods work for critical customer facing operations
- What strategic observations can be drawn as a result of good data quality, for example cross sell?

1.4 Thesis Outline

This work describes the approaches and best practices in the generation, collection and management of data. The rest of this thesis is broken down into following chapters. Chapter two lists the terminology used in this work and the challenges of improving data quality. Chapter three talks about the related work and gives a brief description about various related topics. Chapter four discusses the case study and the findings from it. Chapter five describes the research done and the contributions of this work towards the data quality problem. It discusses the approaches that should be used to improve data quality on a continuous basis. Chapter six explains the proposed framework for data governance- From Data to Knowledge. Chapter seven discusses the effectiveness of this work and finally chapter eight talks about the future work that needs to be done in this area. In the next section we will define further terms that are used extensively throughout this work

Chapter 2: Terminology

We list and define terms that are frequently used in this thesis. This work uses terminology from two areas- technology and business.

Enterprise: It is a business initiative where people work together to achieve a common goal of an organization.

Data: It is a collection of raw facts which can describe the world. Data is of three types:

- Operational data example customer accounts data, inventory data etc.
- Non- operational data like sales data, weather data etc.
- Meta data which is data about data example logical design of the database, database schema etc.

Information: It is the refined data from which some inference can be drawn about the world

Knowledge: Knowledge is retrieved from the information to predict the future in areas like marketing strategy, biological evolution etc

Cross sell: It is an activity where we try to identify existing customers who can be potential customers for selling other products and services offered by an organization. For e.g. a customer owns a home insurance but does not have auto insurance. We can do a cross selling of auto insurance to this customer.

Up sell: This is an activity where an organization tries to sell more add-on for a previously sold product or service to an existing customer. For e.g. A customer has home insurance and his coverage only has insurance against fire and theft. An organization will try to sell this customer additional coverage against storm etc.

Data mining: It is a process to extract knowledge from data using data tools and techniques. Data mining can help a user to analyze data in various ways to better understand the underlying facts about the world. It can be used to categorize the data entities based on a particular property of the data or to identify relationships between them.

Enterprise data warehousing: It is a process to store historical data which is used for reporting purposes. A data warehouse stores data has three functional layers:

- Staging: This stores raw data that is used by developers for testing and other such purposes.

- Integration: This layer is used to integrate all the data and store it in the warehouse providing abstraction for the end user
- Access: This layer provides the end user access to the data in the warehouse.

The best practice is to first identify the purpose of the data in the data warehouse before we begin its design. Most data is used for either customer service or analytics for an organization's marketing strategies. When data is loaded from one or more sources we should understand the level of cleansing and formatting required to transform it according to the user requirements such that it is at a meaningful level of detail for the user to understand and take informed decision based on the data. The Enterprise data warehouse design should be specific to the end user requirements.

Business Policy vs Rule: A business policy is an activity done by top-level management and is a blueprint of the organizational activities. A rule is a very task specific action and it changes at every level and with every operation or task. In this work we have used business policies when we talk about the organizational goals and it is a global reference of organizational activities. Rules are used when we talk about task specific activities like cross selling rules, knowledge extraction rules engine etc. Business policies can effect and change rules but rules do not change business policies.

Extract- Transform- Load: Most data warehouses collect data from various databases and consolidate it into a single database system. ETL is a process of data warehousing that involves the following steps:

- Extract the data from the source systems
- Transform the source system data to match the target system schema
- Load the transformed data on to the target system

One of the major challenges of this process is the correct mapping of attribute- value pair from source system to the target system. The paper "QuickMig – Automatic Schema Matching for Data Migration Projects" by Christian Drumm, Matthias Schmitt, Hong-Hai Do and Erhard Rahm [22] talks about how we can match the schema of the target system with that of the source systems and enhance the mapping of attribute value pairs during the migration of data. Data mapping is usually done manually and is hard to scale when more data sources are added to a system. This causes the problem of wrong mapping of attribute-value pair.

Chapter 3: Related Work

3.1 Enterprise Architecture

Enterprise architecture defines a business structure, its components and their relationships to each other as well as with external entities. It also lays down the guiding principles and standards for the correct functioning and evolution of the enterprise. An enterprise can be any of the following:

- A complete organization- can be public sector or private sector
- A smaller subsection like a single department of an organization
- A component or a system

EA Domains:

Business domain: It defines the business process of an organization to achieve business goals. Business strategies, operations, policies etc are part of business domain which are essential for the correct functioning of an organization.

Information domain: This domain deals with defining the information flow within the organization. It also defines compliance and visibility issues for data that is shared across

multiple touch points in an organization. Data model, database schema design, Master Data Management etc are part of this domain.

Application domain: This domain facilitates the correct implementation for application softwares that are used by end users within and outside an organization.

Technology domain: It defines the softwares, hardwares, middleware etc used by an organization for its smooth functioning and growing customer needs.

Why do we need EA?

In order to improve the efficiency of an organization we need to centralize business processes, make the organizational structure effective and also need quality information. It helps us identify the complexities of the system (a system can be an organization or small units or departments within an organization) and to better align the business processes with IT (Information Technology). For example in our case we are trying to understand an IT system i.e. a data warehouse, which has become highly complex and unmanageable as the organization grows.

In this work we have used TOGAF [32] (The Open Group Architecture Framework) to better understand the system and improve the architecture of an enterprise at all levels i.e. Business, Information, Application and Technology with the main focus on Information domain which will further impact the business policies and processes resulting in changes

and enhancements to the existing applications and addition of new technologies in the organization.

3.2 Lean Thinking

Lean is to maximize customer value and minimize waste at the same time. “Lean means creating more value for customers with fewer resources”. Lean dates back to the time when Toyota was a small company and it could not afford to maintain inventory for raw products nor could it employee a lot of people necessary. Toyota devised a new way for product called “Lean Production” where it invested in a small inventory and decision making was delayed until the last responsible moment. Lean has four basic principles as mentioned in “Lean Thinking” by Mary Poppendieck[11]:

- Add nothing but value: Identify what processes add value to a product or service and eliminate processes which are waste.
- Focus on people who add value.
- Delay until the last possible moment: Do not build up inventory and put revenue on hold because this money can be used for other high priority tasks. Keeping big inventory is a waste in terms of money. Wal-mart practices lean by having a short “shelf-life” for its products.
- Optimize across organization: Apply lean principles not to just one unit of the organization but throughout the organization.

Mary Poppendieck also talks about the types of wastes that exist in various business practices,

- Overproduction
- Inventory
- Motion
- Defects
- Waiting
- Transportation
- Extra processing steps

Our idea of lean with reference to this thesis is to get cleansed data as early as possible by adopting best practices and recommendations mentioned in this work. We try to reinforce best practices and standards on source systems so as to abide by the "Do it right the first time" principle of lean. Instead of loading dirty data and then cleansing it, our aim is to get clean data in the first place. Lean is applied at all levels of this work and not just one. Lean is applied to the process of data generation, data collection, data quality improvement as well as management. In terms of an Enterprise, in order to make the information domain lean, we implement lean processes at business, application and technology domains also. The lean practices are discussed in more detail further in this thesis

3.3 Machine Learning

"Machine learning is a branch of science concerned with the design and development of algorithms which allow computers to evolve behavior based on empirical data such as data from a database or sensors". It is also defined as "Ability of a machine to improve its own performance through the use of a software that employs artificial intelligence techniques to mimic the ways by which humans seem to learn, such as repetition and experience". There are many types of machine learning algorithms[22]:

- *Supervised Learning*: Uses human inputs so that it can map the generated output with the desired output.
- *Unsupervised Learning*: Clustering is a kind of unsupervised learning where data is grouped on the basis of its proximity or closeness of its characteristics to a the other data values.
- *Semi-supervised Leaning*: "It combines both labeled and unlabeled examples to generate an appropriate function or classifier".
- *Reinforcement Learning*: Wikipedia defines reinforcement learning as "Reinforcement learning is concerned with how an *agent* ought to take *actions* in an *environment* so as to maximize some notion of long term *reward*. Reinforcement learning algorithms attempt to find a *policy* that maps *states* of the world to the actions the agent ought to take in those states".

The neural network algorithm we use for our experiment is a supervised learning algorithm. For this purpose we have provided labels to the customer dataset, so that the algorithm can check if the classification it does is correct or not. Machine learning is the most appropriate choice for customer classification because it is a dynamic learning algorithm i.e. with new data quality error types the algorithm adapts itself to perform better by using the technique of back propagation of error to achieve the desired output. We will discuss neural networks in more detail later in this work.

3.4 Data Mining

Data mining is a process of Knowledge discovery. "Data mining is the process of analyzing data from different perspectives and summarizing it into useful information – information that can be used to increase revenue, cuts costs, or both" [1]. IT can help companies to gather customer focused information and take informed decisions related to business processes, market strategies or foretell the market trends. It can be used to determine the impact of one activity on other related activities. Data mining has been used previously to cluster ambiguous data to its nearest neighbor for classification [28]. This paper talks about the effects of dirty data on decision making but the only data quality error they deal with, is missing data values. They try to analyze various market trends and patterns by data mining using data having multiple missing. Data mining is of two types:

- Supervised data mining like logistic regression, decision tree etc

- Unsupervised data mining like clustering, self organized maps etc

According to Wikipedia "In four annual surveys of data miners (2007-2010), data mining practitioners consistently identified that they faced three key challenges more than any others:

- Dirty Data
- Explaining Data Mining to Others
- Unavailability of Data / Difficult Access to Data"

Thus dirty data and unavailability of data can pose challenges to apply data mining to our current problem. However, we have used data mining in this thesis to identify various ways in which data quality is effected for e.g. to identify the impact of data quality on different customer attribute in a particular source system. Data mining did not help us in error identification but it is a very efficient tool to mine knowledge to identify patterns about market demands, target customers etc, once we have data of good quality.

3.5 Balanced Scorecard:

A balanced scorecard explains the business opportunities as a major by-product of why data cleansing and data quality management process is important and how we can achieve it:

Learning: To improve the quality of data, proper training for the CSRs must be made mandatory. A formal introduction of the underlying business processes should be made

for all the stakeholders. We can ensure a high level of commitment from the employee's side, by getting them involved in some kind of incentives or stakes in the business. Also suitable data quality tools should be employed to improve data quality.

Internal: Improve the processes such as the data entry and ETL process to scrub invalid data like special characters, incorrectly parsed data etc to enhance data quality. A good metric to measure the quality of data is to check for redundant customer information and reduced number of duplicates and suspects.

Customer: Increasing business by generating new customers is just one aspect. Customer retention is also very important. Customer loyalty programs etc should are affected due to bad data quality. For example consider a scenario where a customer owns two or more policies and is eligible for a discount on the new policy but due to the multiple customer instance problem we are not able to locate information about the other policies held by the customer. This will reflect bad on the organization's reputation resulting in dissatisfied customer base, causing increased customer attrition rate. A metric to measure this can be customer feedbacks, no. of new customers, customer attrition rates etc.

Financial: As a result of all the above activities, business will benefit in huge ways. It results in an increase in revenue because of new customers and cross selling of products and services to old and new customers etc due to improved marketing strategies and business processes. It also improves the brand recognition for the organization.

This process is not a one time process. In order to maintain high data quality at all times to take correct and informed decisions at all times this process of learning for employees at all organizational levels, internal business policies and procedures and giving high preference to the customers by providing timely service to improve the revenue for organization should never stop.

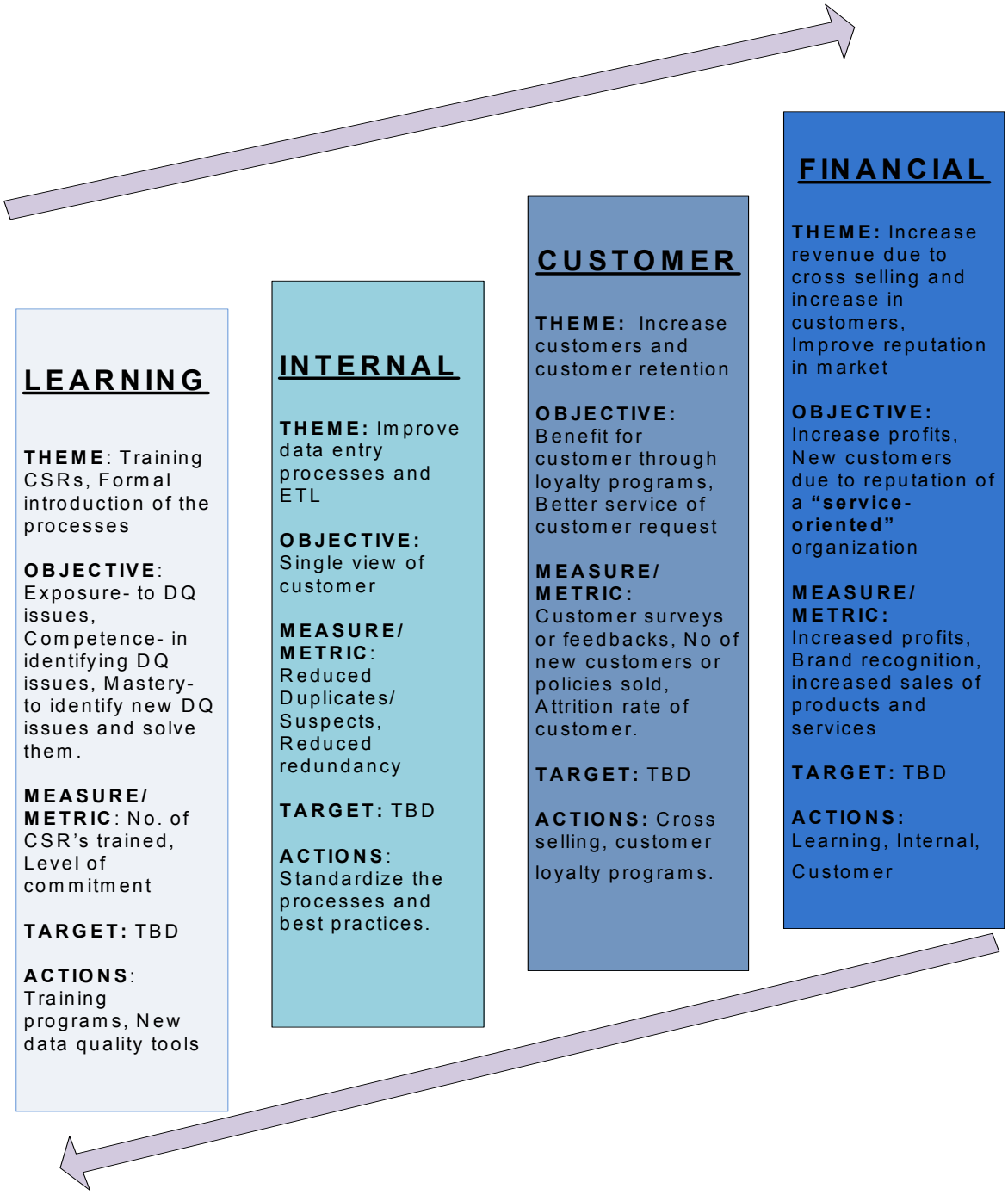

Figure 2. A Balanced Scorecard

Chapter 5: Case Study- Nationwide Insurance

At Nationwide, the data warehouse is accessible for usage by different departments across the enterprise and is shared among many applications and customer interaction job functions. The sharing of the same customer data across various touch points requires consistent definition and format. Compliance and visibility issues are also to be taken care of while sharing data across multiple touch points. ECIF stands for Enterprise Customer Information File and is a central repository to store and manage a "Single View of Customer" data. It is a very important shared asset for the organization because it can enable and provide "One Nationwide" services to its customers. The main issue with data quality is how to keep the data clean while going forward? As time passes by businesses merge and grow. A big organization buys many small organizations and data is collected from all these organizations. This continues over time with businesses buying smaller business and acquiring more database systems. Realizing that there is interoperability of data among multiple applications and departments at Nationwide, and that business rules are not consistent across organization and data and its use is mutable, we need a consistent and standardized view of data. Our main consideration at this point is how do we make sure that we have continuous data quality management processes and what

governance and continuous improvement services should be provided at different organizational levels? The current implementation has the following problems:

- Inconsistency across different source database systems
- The matching rules and parsing rule are inefficient
- Data updates/ changes made to the central repository are not enforced to the source systems
- Bad data quality is causeing:
    - Delayed and unsatisfactory customer services
    - Wrong analytics for marketing and cross selling strategies
    - Dissatisfaction among existing customers

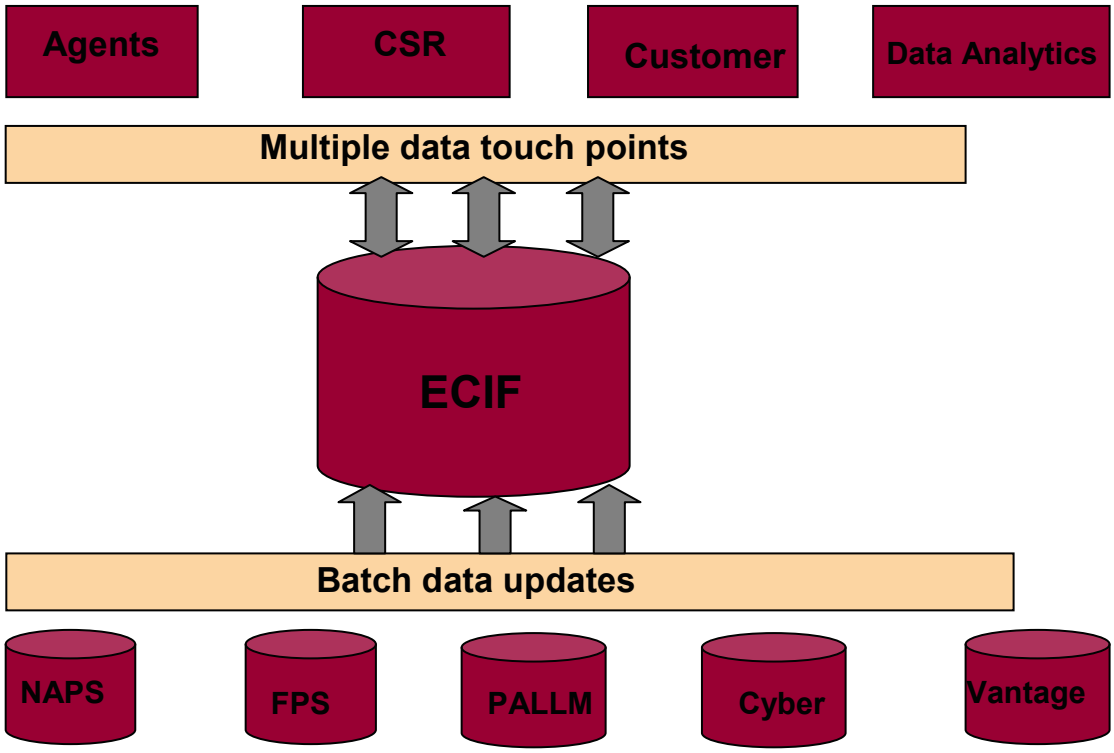

Figure 3. Interactions between the central repository and various data touch points

Data redundancy becomes a huge issue when data is merged from across the enterprise with multiple data sources feeding their data into a central repository. Data duplication causes a big risk to the organization in terms of various resources like time, money, good will etc which results in loss of market and customer base. When data from multiple legacy systems are merged together in a central repository system, multiple instances can be created for the same customer. For instance if System A has an entry for customer Timothy Sweeney and similarly system B also has the same customer Timothy Sweeney with a short name Tim Sweeney, then when the data is collected from these two systems there are two different instances for customer Timothy in the central repository. Also

these two instances for customer Timothy might not have the same data attributes because of the difference in the database schema of every legacy database system. This is a very common scenario in enterprise data warehouse where data is collected from multiple legacy source systems. This can cause chaos in an organization due to unsatisfied customer base and inefficient and wrong business policies due to wrong analytics done on bad data.

5.1 Planning to Achieve High Data Quality

The goal of the planning stage was to achieve the following-

1. A formal definition of what data quality means to the organization: The definition of high data quality is relative to the purpose of its use. Data of one kind might be of high quality for one organization but it might be of poor quality for another organization or department of the same organization
2. How to measure the quality of data: Some user defined function which allows a user to make acceptability decision based on quality data like accuracy, timeliness etc.
3. Identify the interdependence of systems to see what data is shared and how it is being shared.
4. To give a formal definition for data consistency across systems
5. Develop a Lean roadmap

6. Ensure reuse of processes across assets within Customer Information Management

Poor or bad data quality can have severe effects on the overall functioning of any organization which uses this data to analyze market trends and opportunities for improved business. The quality of data depends on the design and production processes involved in the generation of data. To design a better system, offering improved quality data proper planning was very important in order to define data quality in the current perspective. For our purpose high data quality means accurate and timely customer data which can be used across the organization to construct efficient policies and strategies for better marketing and increased sales and business. The below mentioned steps were taken to achieve the planning goals:

1. **The Representation view:** To represent ECIF in user perceivable form in order to retrieve timely information for improving business operations and processes.

2. **The Functional view:** Identification of data quality in terms of specific user needs.

3. **The External view:** It is concerned with the use and effect of the database system. It treats the system as a black box.

- Use: Identification of who the end user of the system is and for what purposes the system will be used.
- Effect: To identify and analyze the relevance of the system and support its existence?

4. **The Internal view:** It is concerned with the operation and construction of the system. It's the white box view of the systems interiors.
    - Operation: Study the necessary operations to attain the required functionality of the system. Processes such as data entry, data maintenance, data delivery etc are identified and analyzed for existing defects.
    - Construction: To design and implement the system to achieve the end-user requirements.

5.2 The Process of On- Going Quality Problem

The data quality problem analysis started with a descriptive research to identify the existing data quality error types in the provided database and also to identify the root cause of those errors. The problem analysis also involved to analyze the impact of each kind of problem with a measure of it being high, medium or low. During the research I identified multiple error types caused due to the wrong practices followed during the generation and collection of data. These errors are caused essentially due to three different interactions throughout the generation and collection of data: human-system interactions e.g. when a human enters a data value into the system through some user

interface, human- human interaction e.g. when a customer calls up a customer service center and talks to a customer service representative to enter its information into the system or system- system interaction e.g. when data is transferred from source database systems to a target database system.

5.3 Corrective Actions

Based on the analysis of the data, I provided recommendations for quality check before and after the ETL process so that waste practices are eliminated and data is cleansed before it is been loaded to the central repository. These recommendations save a lot of time wasted in bad practices of storing dirty data and cleansing it, saves time during querying customer information. It also saves a lot of database space which was previously used up by duplicated erroneous data. Data quality improvement and management will also help save a lot of money and generate more revenue for the organization by more cross sell, customer retention and well defined and efficient strategies for marketing. These recommendations have an essence of Lean practice where our goal is to get good quality data as early as possible. To insure continuous data quality improvement we should check level of data quality from time to time on a quarterly or yearly basis. *Leo L. Pipino, Yang W. Lee, and Richard Y. Wang [5]* talks about two assessment approaches for data quality evaluation in their paper **"Data Quality Assessment".** During these data quality reviews two major kinds of assessments should be done- Objective assessment and subjective assessment of data quality should be carried out. An objective evaluation of data is done by maintaining a matrix with a record

of existing data quality issues and updating it every time to incorporate new data quality issues that have come up or in case if the earlier data quality issues have reduced or increased, in order to assess the current level of data quality. A subjective evaluation of data is done by talking to the end users of the data and evaluate the quality of existing data considering the level of satisfaction of the user which can help us evaluate whether the data is accurate, timely and complete or not.

5.4 Neural Network based Customer Classification

In order to reduce duplicated customer instances we need to classify these instances whether they belong to the same customer or not and then merge these together to form ECIF (Enterprise Customer Information File) as a "Single View of Customer". To help solve the problem of customer classification we can use a feed forward neural network based classifier provided by MATLAB. This helps us to classify whether two or more customer instances are same or different. We have tested datasets with different data quality issues and compared their results later during the process.

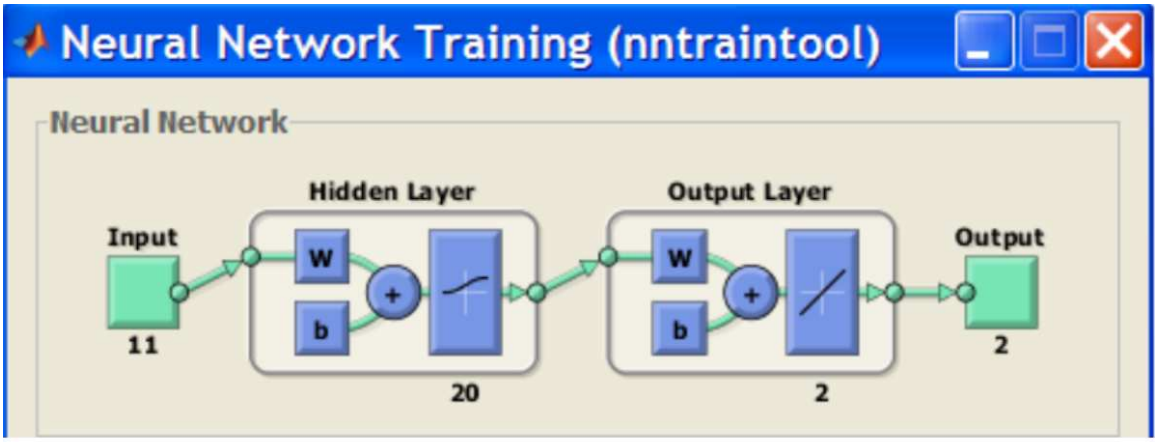

Figure 4. A Neural Network [22] with one hidden layer and 20 neurons

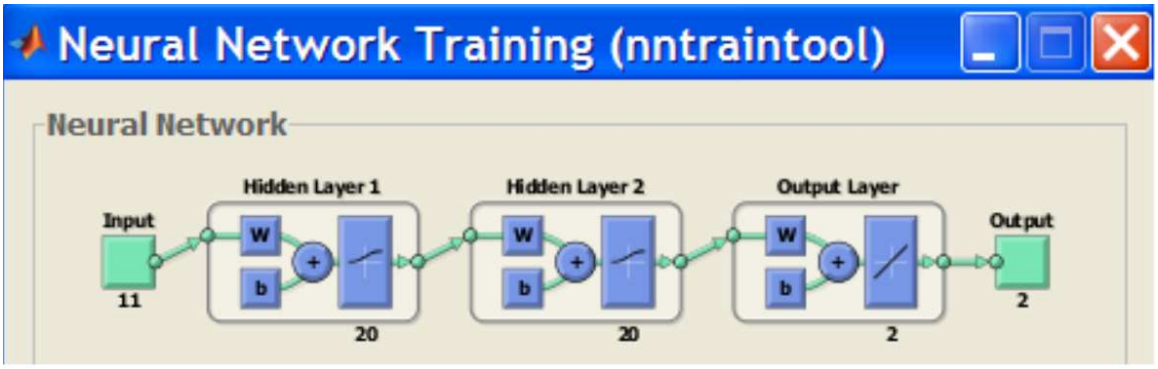

Figure 5. A Neural Network [22] with two hidden layer and 20 neurons per layer

A rule based system is currently in use at Nationwide for customer classification. This system uses a set of rules to identify two or more instances which belong to the same

customer. An example of one such rule can be if two customer instance has same first name, address, date of birth and SSN, they belong to the same customer and the system will merge these instances together into a single customer. The database still has multiple duplicated instances for the same customer because these customer attributes can have various data quality errors. We need services that can dynamically change its “rules” for new data quality error types occur in the data. For this purpose a neural network based machine learner is the appropriate choice.

A feed forward neural network is one which is connected while moving forward from inputs to the output. This network uses the most common propagation technique called the backward propagation where the error flows backward each time the network makes a mistake in the output value.

A neural network is an artificial system which tries to mimic the properties of a biological neural network by using artificial neurons and interconnecting them. It can have one or more input layers which can be the original data or input from the previous layer. Output from each previous layer becomes the input for the next layer. The hidden layers can have multiple neurons which resemble the biological neurons, as they perform the same function but artificially. The inputs are provided some weight and the weighted sum of the inputs are fed to the neuron of the next layer. Each neuron has a threshold value which is then subtracted from the weighted sum of the inputs to that neuron forming a value called the activation of the neuron. The activation is fed to the activation function also called a transfer function. A transfer function can be a logistic function

usually log sigmoid which is a “S” shaped graph or a purelin function which is a linear graph.

Sigmoid function errf(x) = y= 1/ (1+e $^{-x}$) where x is the input to a neuron. It can only take output values in the range 0-1. As x goes to positive infinity y reaches 1 and when x goes to negative infinity y reaches 0. While a purelin function can take output values in a larger range.

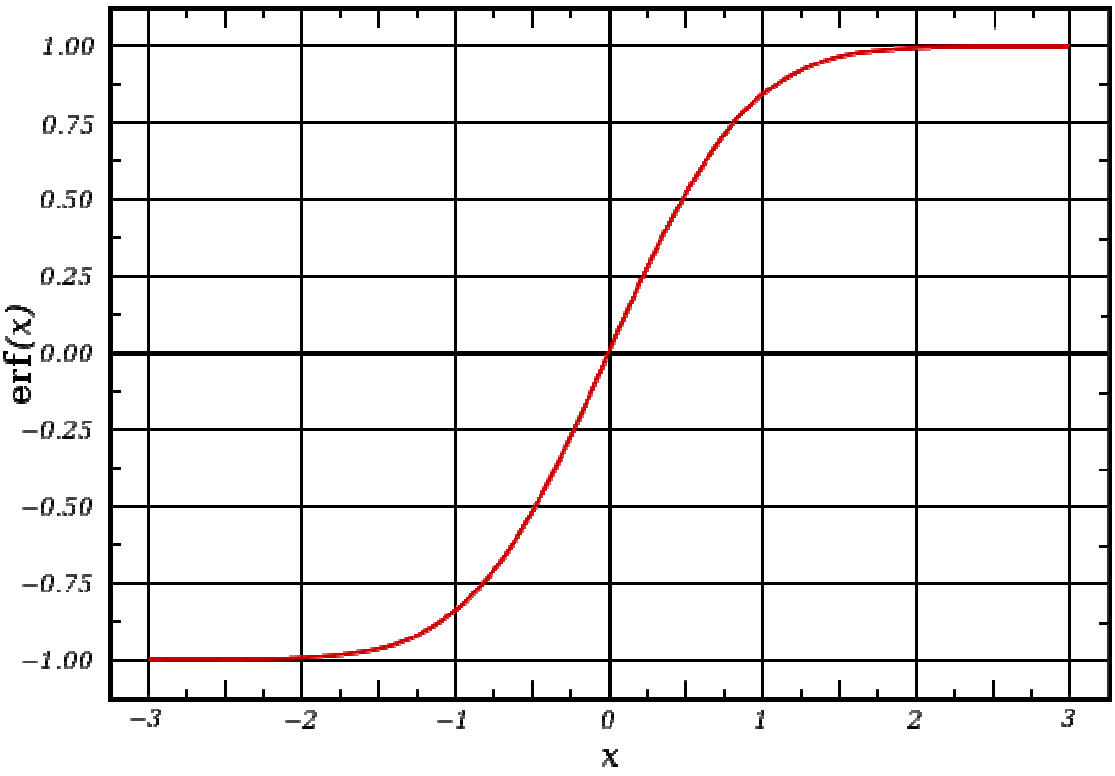

Figure 6. *A Sigmoid function

* Figure taken from Wikipedia

Chapter 6: Research and Contributions

6.1 Problem analysis, classification and quantization

6.3.1 Data Collection

For this problem we collected error data (i.e. data set with existing error types in it) at two stages- data coming from the source system before ETL processing and data being loaded to the target system after the ETL processing. The data was a set of millions of customer records and it was provided by Nationwide Mutual Insurance. It was customer data with customer information like, first name, last name, address, date of birth, marital status, suffix, prefix etc. We then performed adhoc manual data analysis on the data to identify various data quality error types and their sources which primarily fall under these four classes of data error types:

- *Inconsistent*: The data is inconsistent across different legacy systems. For e..g system A has first name, last name while system B has fName, LName etc.
- *Incomplete*: The database has missing values for customer attributes. Important customer information like first name, phone numbers etc are not in the database. Incomplete data is generated due to blank or missing values entered by, party using online system to fill fo

    - When one system with a different data model feeds another system.

- *Garbled*: When real world information is not of utter importance for a particular system and any value entered in that field will suffice, it creates meaningless data states. These are particularly caused due to absence of quality checks during data entry. For e.g. Contact number of a customer is 2020202020 or his SSN is 01234.
- *Ambiguous*: This is caused when one real world state (e.g. a single party is duplicated as two different parties) is mapped to two or more database states. For e.g. one customer instance has first name as Timothy and the other one has first name as Tim. This causes ambiguity in the database.

6.1.2 Existing Error Types

S. No.	Error Type	Root Cause	Source system	Recommendations	ETL Work flow section	No. of records affected	Impact
1	Truncation	Field/ Attribute size (sent by source system)	PALLM	Standardize the attributes field length on the Front end UI		Few hundred	Medium
2	Truncation	ETL Processing (smaller field size in target database)	PALLM	Field length in target database greater >= to source system attribute field length	Work flow 7 & 8	Few hundred thousand	Low
3	ETL	Wrong Parsing	PALLM	ETL code should perform common translations and transformations	Work flow 7 & 8	Millions of records	High
4	ETL	Mapping errors	PALLM	Target attribute field names should be similar to source system field names	Database	Few thousand	Medium

Table 1. List of existing error types **Continued…**

Table 1 continued….

5	ETL	Null or values like Special characters(#,%*/ etc)	PALLM	Data scrubbing	Work flow 5, 7 and 8	Thousands of records	High
6	Typos	Erroneous entry (unavoidable essentially by CSR)	PALLM	Better CSR training		Few but not negligible	Low
7	Invalid data	Invalid data sent by source system	PALLM	Data validation checks like real-time checks put on the name field to see if a number is appended to first or last name or for SSN verification online		Thousands of records	High

This table lists the existing data quality issues in the ECIF database provided for this study by Nationwide Insurance. The database is essentially a customer database with various customer attributes like first name, last name, DOB etc. A detailed impact analysis is done of various data quality issues using data mining and getting database counts to categorize them as high impact issues, medium impact issues or low impact issues. A huge number of issues are caused due to wrong parsing of data during ETL which causes wrong mappings of attribute- value pair from source system to target database.

6.2 Root cause analysis

A manual analysis of data within ECIF suggests that the quality of data is very low. There is existence of data quality errors where exists:

- Invalid contact numbers
- Wrongly parsed First and Last names
- Name field with date/ special characters/ Account numbers etc
- Dangling duplicates and suspects in PERSONNAME table.
- Parties with no names
- Invalid SSN

The sources of these errors are due to various operational and design failures and can be broadly classified into three categories- Human- Human interactions such as interaction between a customer and a customer service representative, Human- system interactions such as customer and the front end UI and System- System interactions such as front end UI and the database. There were broadly four data error types-

- **Truncation errors**: There are essentially two causes for truncation error. One, the field length of the source system UI is small and the customer service representative truncates the name field while doing data entry.

  Two, the field length of target system is small and the name field is truncated during ETL while data is loaded from source system to the target system.

- **ETL errors**: The causes for Extract- Transform- Load errors are wrong parsing which causes mapping errors of attribute value pairs from source

system to target system account number of DOB to be mapped to first name etc., inefficient parsing logic which adds the prefix or suffix of a customer to its first name or middle name.

- **Typos**: Typos are also caused due to human- system interaction where a human enters Alice as Slice or 12345 as 123456. Erroneous data entry can be caused due to: Customer service representative did not understand what the customer said over the telephone, lack of attention and motivation etc.
- **Invalid data**: Invalid data is sent by source systems itself because agents or humans enter invalid data into the required fields if they do not have that data with them. This invalid data gets saved to the database as there are no validations to check if the data entered is correct or incorrect. There is also a huge amount of invalid data like phone numbers of type 20202020 or SSN of type 123456789 which is stored during human-system.

which are further caused due to various sources like wrong mapping of attribute- value pairs, incorrect parsing due to wrong parsing logic during the ETL process, typing errors during data entry, truncation due to small field length and large spelling of names and addresses etc.

ETL processing is causing a huge chunk of dirty data in our current scenario. The extract load and transform process is causing various data quality issues caused mainly due to three reasons. One reason for dirty data generation due to ETL process is the wrong mapping of attribute – value pairs. During the ETL process data comes in various formats,

flat files, excel files etc and the database schema also changes with every source system. A name is stored in system A as FName and Lname while in system B it is stored as First_Name, Given_name_two, Given_name_three and Last_name. During the ETL processing this can cause wrong mapping of first name and last name and middle name in different fields of the target system. Similarly, a person's account number is mapped to his first name and his date of birth is mapped to his last name. ETL process also causes wrong parsing which results in another data quality error type. Due to wrong parsing logic a person's name is incorrectly parsed and his suffix is stored with his first name creating a new but duplicated customer in the target system. ETL also causes truncation of names and other fields if the size of an attribute in the source system is larger than the size of the attribute field in the target system e.g. customer named CECILIA is stored as CECIL.

6.3 Data Mining- Usefulness and Limits

In addition to the manual analysis, we also used STARLIGHT a data mining tool to identify and prioritize for fixing the existing data quality issues. These Figures are generated by the starlight tool which help us better understand the problem by providing visual results to categorize data quality issues based on system types, attributes, stakeholders like type of agents selling products and services or last update users who made changes to the data which has errors. In the first Figure it is very clear that the first name field has the most number of errors while address field has least number of errors.

In the second Figure we can see that the products and services sold by Independent channel agent have the highest number of data quality issues while products or services sold by Employee agents have minimum number of data quality errors. Starlight did not provide any additional insight into the problem. It helped us to better visualize the different ways in which data quality was affected. For example we tried to visualize which customer attribute e.g. first name, last name etc. was getting impacted the most and which source system was sending the most dirty data. Additional insights about which employee is creating which kind of dirty data etc could be obtained if we had additional data for this purpose. We also ran the mining to get results as to which stakeholder was creating the biggest chunk of dirty data.

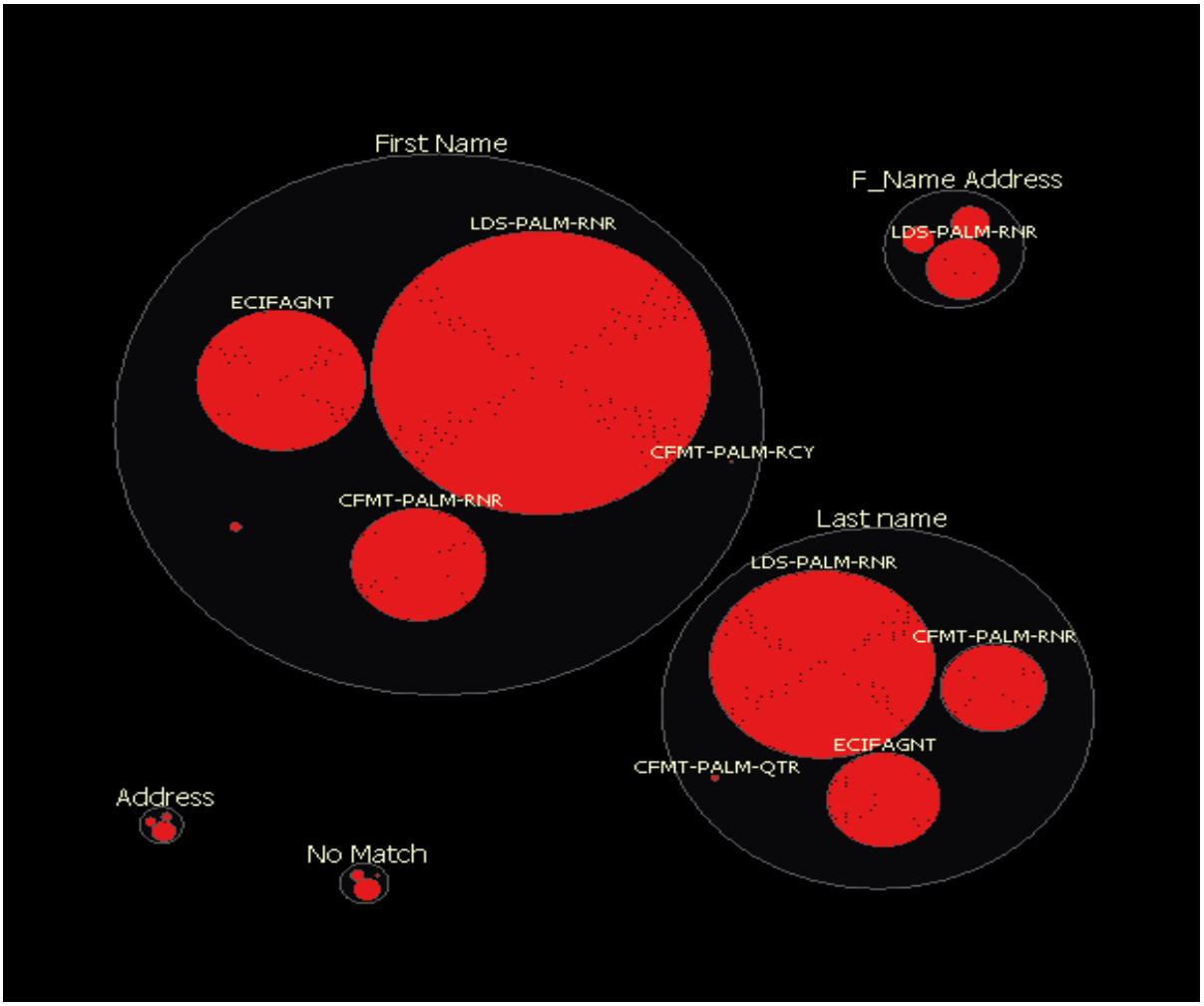

Figure 7. Starlight Generated Figure for error classification based on customer attributes and source system generating the data

Figure 6 depicts the chunks of error data affected due to corrupted customer attributes like First name, Last name etc and the systems generating that data e.g CFMT-PALM-RNR etc

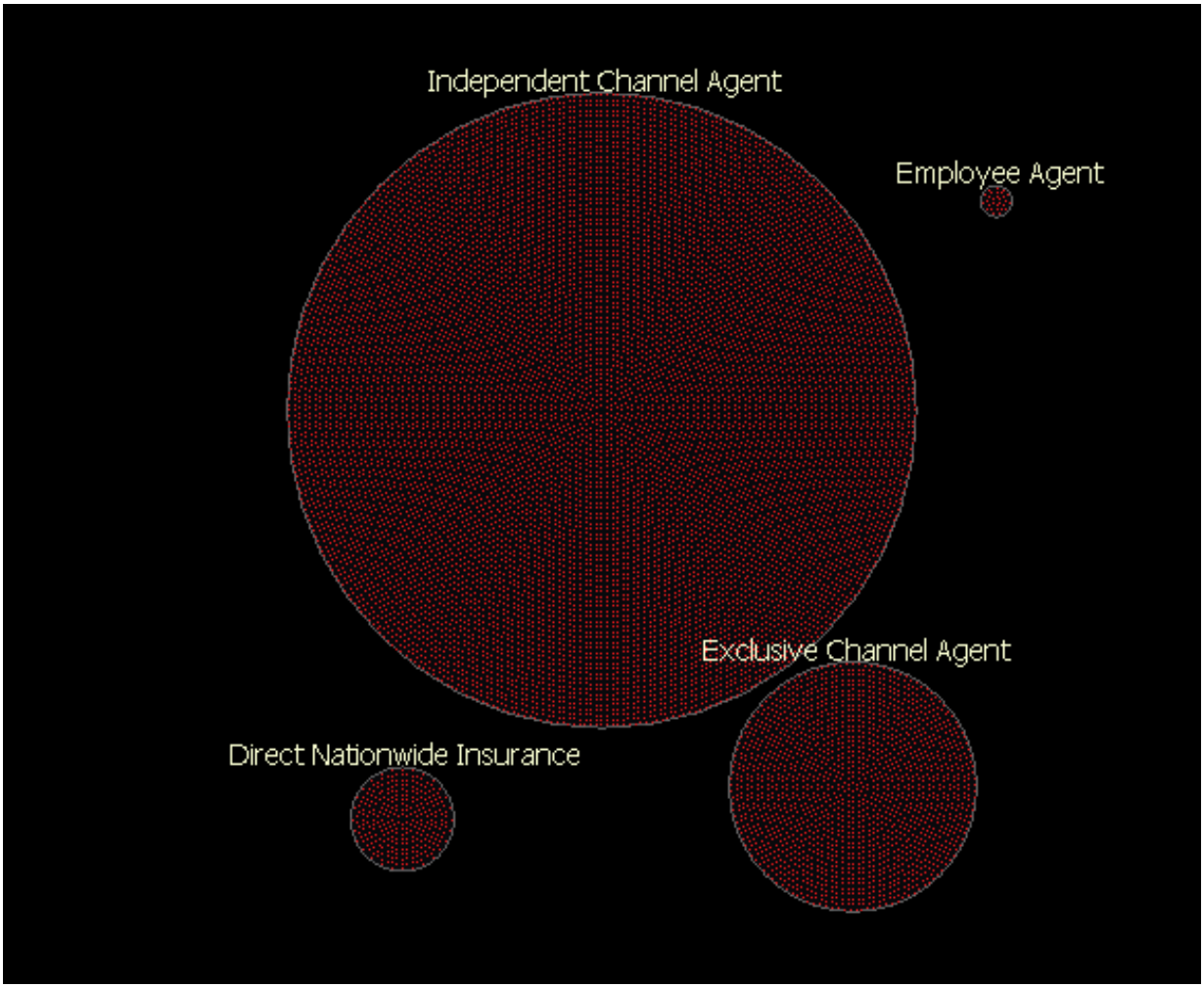

Figure 8. Starlight Generated Figure for error classification based on stakeholder generating the data

Figure 7 depicts the error data chunks created by different types of agents and stakeholders involved in the generation of data. Starlight is a good visualization tool that gives us a better understanding of the data quality defects but does not provide any additional information. Starlight is limited in its application due to its inbuilt rules for data mining which can not be customized for user specific task. Also the kind of data used for a specific task also limits the usefulness of starlight. Absence of a data attributes

can limit the application of Starlight to any project. For instance we could not use Starlight to mine information about how many errors were caused due to truncation or due to wrong mapping of attribute- value pair because there was no explicit information about these issues in the dataset itself. Thus, Starlight is limited in its application and use across projects and tasks.

6.4 Recommendations for Data quality Improvement

We provide two different types of recommendations to improve data quality in an on going fashion. We are trying to make the data collection and generation process lean by making changes to the business processes and application implementations and getting clean data as early as possible.

Process improvement: These recommendations suggested as a precautionary method to ensure that we get high quality data at the initial stages like during data entry etc. This is the first layer of data quality check, when data is entered from an application interface into the source systems.

- **Training:** An introductory training should be provided to each employee so that they understand the importance of data and accept it as their responsibility for data quality improvement and management. Also provide proper training to customer service representatives and the developers who work in the development of the database as well as the

development of the Extract-Transform- Load process codes. They should follow best practices and standardize processes to achieve high data quality. Exhaustive testing should be done before the data or the underlying processes are put to production.

- **Online validation of data:** Services should be implemented to validate data value at the time of data entry. For e.g. a validation service should check the value of a name field, so when a number is appended to a name field, it should display an error so that the user checks and corrects the value.

- **Online verification of data:** A verification service should be implemented to verify data values during data entry. Example a person's SSN should be verified which will connect to the SSN site online and verifies if the SSN entered is correct or not.

- **Standardize**: Processes for generation and collection of data and database schemas should be standardized throughout the organization. For e.g. standardize attribute values like DOB as mm/dd/yyyy instead of mm/dd/yy , dd/mm/yy. Mm/dd/yyyy etc

- **Synchronize**: The data in the source legacy systems should be synchronized with the data in the target system everytime changes are

made to the data in the central repository. Timestamps should be provided to the data in the target repository and source systems to ensure effective synchronization

When the cost of implementing process improvement measures becomes higher than the value earned due to good quality data, it becomes meaningless to apply process improvement changes. For e.g. changing the code to standardize the database schema of a legacy source system is very costly as the code can be millions of lines long and most probably the people who wrote that code have already left the organization. In such scenarios, instead of making changes to the source system processes we should take corrective measures to improve data quality.

Corrective measures: Corrective measures are used to capture bad data if it has entered the system due to some operational failure or design failures. We implement processes for revalidation and verification of data during the migration of data from source system to the target system which enables a second layer for data quality check.

- **ETL Verification of Data:** Contact numbers with just 3 or 4 numbers, with alphabetic characters, or 202020202 etc should be deleted from the database. Proper checks should be implemented during the Extract Transform and Load process to identify invalid contact numbers and send an error log to the source system and get it fixed.

- **ETL Verification of Data:** SSN verification check should also be implemented during the ETL processing so that the wrong SSN's not captured during the online data entry will be captured during the ETL and a log will be sent to the source system which will then contact the customer or the agent who sold the product to update the SSN to its correct value E.g throw error for SSNs like 101010101, 000000000 etc.

- **Data scrubbing**: The ETL code should be enhanced to detect and delete special and invalid characters, dates, numbers etc from name fields or characters like %; * etc from various data fields like names, phone numbers, address etc.

- **Improved processes for capturing and collapsing Duplicates/Suspects**: The duplicate and suspect identification and collapsing logic needs to be improved for better identification of same customer instances. There are multiple instances for the same customer existing in the target system even after the Duplicate Suspect Processing has been done on it. Using a neural network for customer classification should be used for this purpose.

6.5 Classification of customers using Machine Learning

Over time new systems are acquired by an organization. This results in new type of data coming in with different data attributes and new information related to customers,

products and services. A rule based system for customer classification has static rules which are not updated each time new information flows in. However a machine learner acts dynamically as it tries to learn new data quality error types and adapts its "rules" according to the changes in data. In the paper "Identifying Relevant Data for a Biological Database: Handcrafted Rules Versus Machine Learning" by Aditya Kumar Sehgal, Sanmay Das, Keith Noto, Milton H. Saier, Jr. and Charles Elkan the experiment shows that the machine learning algorithm performs better than a static rule based system because the machine learning algorithm tries to learn every time it makes an error, which makes the classification dynamic. Use of neural networks for customer classification is also discussed in another doctoral dissertation "**Solving the data duplication problem for complex databases using neural networks**" by Abdullah Abdulrahman Al-Namlah[30].

For our experimental setup we created a dataset of 1000 records for 201 unique customers. The dataset consisted of eleven different customer attributes – Suffix, First name,Middle name, Last name, Address line one, Address line two, City, Zip code, DOB, Gender, Phone number. For the purpose of classification we add labels to this data to inform the classifier whether two or more instances match or not. This label is manually added as the 12^{th} attribute in the dataset which helps the algorithm to check its results. We create four different data formats- binary distance data, ASCII distance data, Levenshtein distance data and normalize Levenshtein data which we feed to our neural network because our dataset has string values and a neural network works best with numeric values. There were the following error types in our dataset:

- Address non-match for same customer (15%)
- Multiple non-matching telephone numbers for a single customer (10 %)
- Truncated or missing attribute values (2%)
- Date of birth off by a digit (1%)

We use a Crab Classifier provided by MATLAB. This classifier is a very close fit for our dataset as it is used to classify a crab as male or female based on six different crab attributes which matches our customer dataset of eleven different attributes. The inputs for our neural network classifier were 11 customer attributes and the final output was either a match or a non match for customers. The stopping criterion for our algorithm is such that, when there is no remarkable improvement in the output of the classifier for training and validation data, it stops.

6.5.1 Splitting Data for the Experiment

The dataset is divided into three parts.

- **Training set:** This data is used to train a neural network
- **Test set:** It is used to evaluate the results by measuring the number of pairs the algorithm correctly classified divided by the total number of pairs in the testing set.
- **Validation set:** The validation set will be used as a stopping criteria to ensure that over-fitting does not occur

The default setting of the network divides the data as Training data, Test data and Validation data.

6.5.2 Feature Vector Extraction

Since the mismatch in customer instances in the ECIF data was primarily due to character differences caused by typos, truncation or short names, distance feature extraction seemed to be the obvious choice. The feature vector used by the neural network is the difference between the attributes in the two instances. Each instance consists of eleven customer attributes. We tested four methods of calculating the difference between the attributes: binary distance, ASCII character distance, Levenshtein distance and Levenshtein ratio.

6.5.2.1 Binary Distance

The binary distance is the simplest method. If the values of an attribute in the two instances are identical then their distance is 0, otherwise it is 1.

6.5.2.2 Euclidean Distance

ASCII character distance is the square Euclidean distance between two attribute strings. 4

distASCII (S1, S2)= $\mathbf{N}\ \Sigma\ \mathbf{I}$ (S1,I +S2,i)/ 2

where S1 and S2 are the attribute strings, N is the number of characters in the strings, and Sx,I is the *i*th character in string Sx. Since the attributes are not guaranteed (sp) to be of

the same length, the strings are resized to the larger of the two lengths with " " used to complete the smaller string.

6.5.2.3 Levenshtein Distance

Levenshtein distance is the number of changes that a string undergoes so that it becomes identical to another string. For example, the Levenshtein distance between "Saturday" and "Sunday" — "a" and "t" are dropped and the "r" changed to an "n".

The Levenshtein ratio is the normalized Levenshtein distance

ratio= 2 * distLevenshtein (S1 , S2)/(|S1| + |S2|)

6.5.3 Experimental setup:

The execution of the experiment has six primary steps:

1. Create the network for customer classification
2. Configure the network.
    - One hidden layer or two hidden layer
    - 20 Neurons per Layer
    - Logistic Sigmoid (logsig) or Linear (purelin) Output Activation Function
3. Initialize the weights and biases by assigning random values
4. Train the network using pre labeled data which informs the algorithm about its classification. In our experiment we have labeled the data such that all customer instances which belong to the same customer have the same identification number.

5. Validate the network to confirm the results of the trained network by providing new data which is similar to the training data to avoid over fitting or under fitting.

6. Use the network for classification

6.5.4 Results and Conclusion for Customer Classification

Neural networks have been widely used as proficient classifiers. It gave some very good results with our dataset. Dataset obtained from Nationwide Insurance's customer database was tested with different settings of neural network, to evaluate the performance of the neural net toolkit on the customer classification problem. Although the data had a lot of noise i.e. missing valus, incorrect values etc, our problem of predicting a customer instance as Matching or Non- Matching was a fairly easy one for neural networks after we converted our dataset into numeric datasets. Neural networks make estimates to predict the correct output i.e. it reconstructs missing values with the information provided in the data. This estimate can be made by replacing the missing data values for an attribute *A* with either a zero, some other random value or a mean of all the values of *A* in the training dataset. This process of classification based on missing data is also called imputation. For our dataset we have already replaced the missing values by a '0' in all datasets. Also incorrect data values are assigned a numeric value which is tested each time to see if it matches the pattern to fall into one of the two classification classes i.e. Match or Non-Match. If it doesn't match the patter, the network makes changes to its network using back propagation technique and tries to make a correct approximation each time.

The binary distance test set gives 99.6% correct classification for our problem. However, ASCII distance dataset also performed well with a correct classification of 98%. This was a small dataset so the results might go down a little with a large dataset. Using a validation set for early stopping also helped in addressing the problem of over-fitting. The performance of binary set with a purelin transfer function and two hidden layer is poor as compared to one hidden layer. A possible reason for this is that the unnecessary addition of hidden layers to the network takes up more training time. This causes the network to memorize the patterns on the training set and become specific to the training data. But when this network is run on the test set it is not able to do correct classification on the new data and hence the performance goes down.

Input	**Binary**			
Function	**Logsig**		**Purelin**	
Layers	**1**	**2**	**1**	**2**
% Correct	98.7	99.6	99.3	98.3

Table 2. Results for Binary distance dataset

Input	**Euclidean**			
Function	**Logsig**		**Purelin**	
Layers	**1**	**2**	**1**	**2**
% Correct	44	98	97	97

Table 3. Results for Euclidean distance dataset

The ASCII distance performs better with two hidden layers and log sigmoid function giving 98% correct classification. With a purelin function the classifier gives 97% correct classification with both one and two hidden layer neural network.

Input	**Levenshtein**				**Normalized Levenshtein**
Function	**Logsig**		**Purelin**		**Logsig & purelin**
Layers	**1**	**2**	**1**	**2**	**1 & 2**
% correct	64	44.7	84.7	94.7	96.3

Table 4. Results for Levenshtein distance dataset

Levenshtein distance does not perform very good in customer classification. With the network configuration using log sigmoid function it gives 64% correct classification

using one hidden layer and 44.7% correct classification using two hidden layers. The performance gets a little better with a a purelin function giving 84.7% correct classification using one hidden layer neural network and 94.7 % correct classification using two hidden layers. The performance of normalized Levenshtein distance data does not change with either purelin or log sigmoid function using one or two hidden layers for the network. In all the cases it gives a correct classification of 96.3%.

Chapter 7: Governance related to Information and Knowledge

The adaptive framework for process governance is based on the ACE architecture. The ACE architecture was proposed and developed by Dr. Jay Ramanathan and Dr. Rajiv Ramnath at CETI, The Ohio State University [29]. ACE is the Adaptive Complex Environment which has three primary building blocks, namely, Roles, Assets and Interactions. An interaction can be defined as a transaction between a customer and an Agent (also called Role). A simple example of an interaction is when a customer calls the Customer Service for a certain enquiry related to a product or service of the customer's interest. An event triggers an interaction. This event can be a request for a certain product. A role is an Agent who provides services to complete an interaction. In our example the customer service representative is a role. An asset is produced when the interaction completes. For example, A customer initiates an interaction when he calls the customer service representative who plays the service provider role. The customer buys a product and his data is stored in the company's database. This customer information is now becomes the asset.

A goal can be described as stakeholder objective achieved through a related set of steps. Goal modeling is used to describe processes for software development, requirements gathering, business objects etc[31] There are different goals for different tasks and at

different organizational levels. A goal at lower level can be a subset of the goals at levels above it. This means that the attainment of a lower level goal does not imply the successful attainment of the higher level goals. However attainment of a lower level goal contributes to the attainment of higher level goals. For e.g. good data quality can help achieve better and effective marketing strategies and increase business revenue. But there are other factors which also contribute to completely achieve the ultimate business goals like business policies for customer loyalty programs, efficient workforce etc.

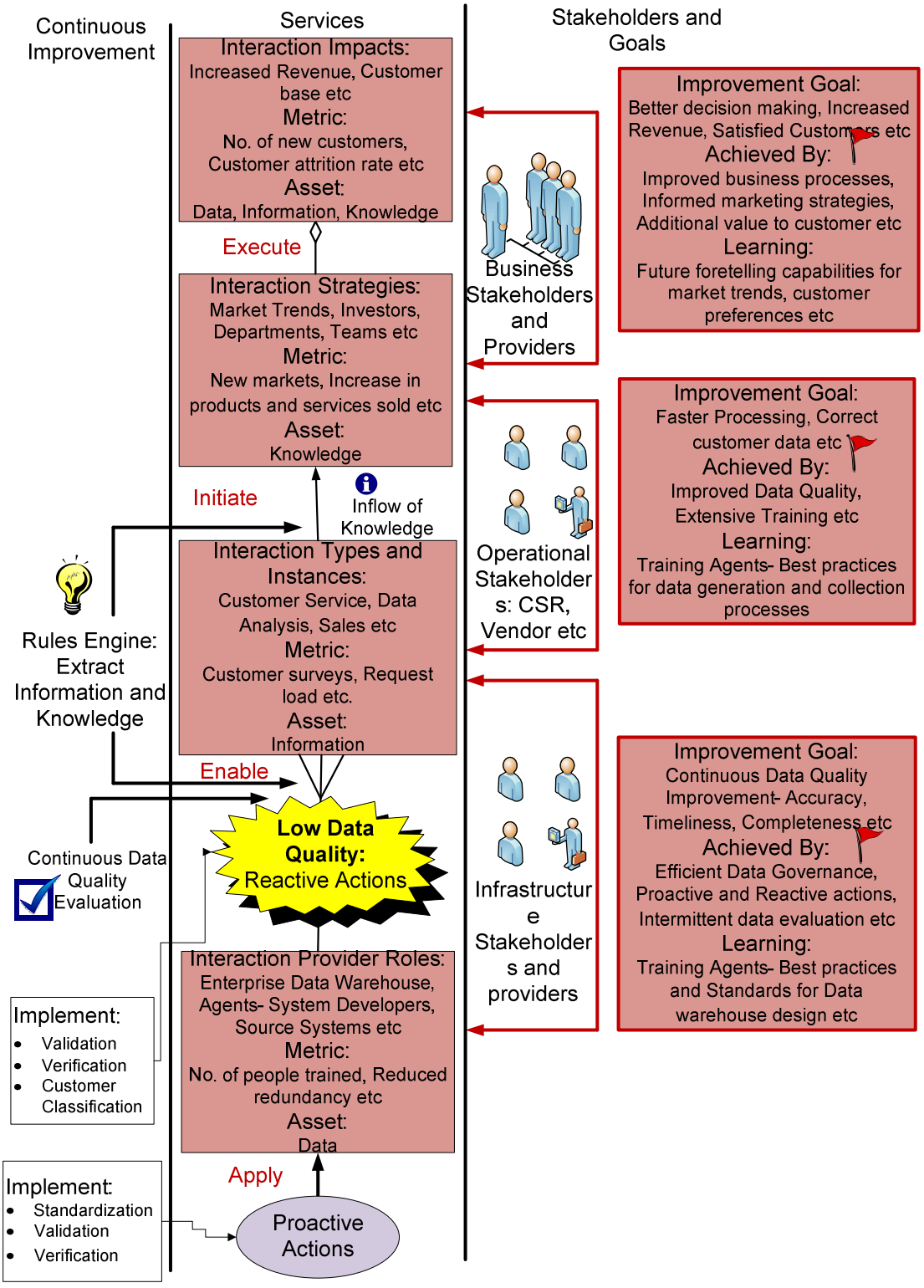

Figure 9. An Enterprise Framework for Process Governance

An ACE framework divides the organization into four operational levels- Infrastructure, Operations, Strategies and Business. We use this framework for our work, in order to make it customizable for process governance for various organizations. There are multiple stakeholders involved in an organization. There are three types of services provided at different organizational levels using the adaptive framework for process governance:

- **Infrastructure Services**: As Master data management servers and data warehouses are used for decision making, their correctness is imperative to the organization, to avoid incorrect conclusions. The value of data assets appreciates with the potential of new and improved uses of data. Quality of data is of utmost importance demanding more sophisticated data quality tools and technologies for continuous data quality improvement and management. The stakeholders at this level are EDW, software developers, source systems etc. The infrastructure level services aim at continuous data quality improvement in an organization that uses enterprise data warehousing for customer information management, customer interaction management etc.

  This is achieved by having an efficient data governance team which takes responsibility for data quality management and is the primary contact in case of any issues related to data quality. They should reinforce and standardize the practices for creation of data warehouse, taking process improvement and corrective measures for continuous improve data quality etc. An efficient

metric to measure the impact of these services is to see the number of people getting trained, Reduction in redundant data etc

- **Operational Services**: If the accuracy, availability, or timeliness of the data a company uses or produces is compromised or in doubt, the value of the data diminishes. This can result in non optimal operations which can delay or misdirect corporate initiatives within the organization further causing personal and customer dissatisfaction and frustration. This can pose huge financial and reputational risk for the organization. With improved and cleansed data the operations for executing business policies will become more efficient. The stakeholders at operational level are customer service representatives, vendors etc. The operational level services aim at handling higher customer request load, maintaining correct customer information etc. Information retrieved from the data collected at the infrastructure level is delivered to these agents to perform tasks better and achieve business goals.

  This can be achieved by having good quality data which is a by- product of successful implementation of infrastructure services, extensive training for the provider roles etc. The metric to measure the effectiveness of operational service can be by conducting customer satisfaction surveys, ability to handle more customer requests etc.

- **Strategic Services**: In order to grow, prosper and maintain a competitive edge in business, a company needs more information which is equivalent to more data of

better quality and variety. If we do not have the right data with accurate and complete information related to every customer, it can result in chaos within the organization and the market due to incorrect customer leads and prospect generation or wrong identification of new markets etc. Knowledge extracted from the information at operational level enables informed decision making at strategic level. Stakeholders at the strategic level are investors, departments, teams etc. The services at the strategic level aim at achieving better decision making powers, increased revenue and having a satisfied customer base etc.

These service levels can be achieved by having effective business policies, improved marketing strategies or by providing additional value to the customer etc. A metric to measure the effects of strategic services is to see if the number of products and services sold has increased, whether new markets and customer preferences have been identified or not etc.

The red arrows in the figure, indicates the two way flow of feedback from the level above to the level below it. That is, the infrastructure design is affected by the operations of an organization; the operations affect the strategic decisions within the organization which further affects the business policies. This is a cause and effect cycle where the business policies can cause a change in the strategies which can cause changes at the operational level. For e.g. the purpose of data in the database effects the database design (infrastructure level) and the purpose of the data is actually the operations performed on that data (operational level). Example of these operations can be updates on customer data or knowledge mining for marketing analytics.

The learning at each organizational level differs. Learning at the infrastructure level deals with training the developers, agents, employees to follow best practices and standards for the creation of an enterprise data warehouse. The data warehouse design should be driven by end user requirements in mind i.e. the purpose for which the data in the warehouse will be used. The asset at this level is factual data about the customer. The learning criterion at operational level is to train agents like customer service representatives, vendors etc to follow standardized practices in the generation and collection of data. This training ensures that we get clean data as early as possible, because if garbage gets into the system, garbage comes out of the system. The data at the infrastructure level is now used as information of a customer which is the asset at the operational level. The learning at strategic level empowers the organization with future foretelling capabilities for market trends and competition, customer preferences in order to provide more value to the customer and generate more revenue for the business. This learning is achieved by using the knowledge extraction from the data at the infrastructure level. This knowledge is the asset at the strategic level.

Chapter 8: Effectiveness

Neural Networks for customer classification: The approach for fixing the problem of customer classification using neural network is very efficient and will fix the problem of redundant customer information by 100%. This also enables continuous data quality management of duplicated customer information as neural networks dynamically learn when new error data comes into the system by making changes, every time the network does a wrong classification.

Process improvement and corrective actions: Also the recommendations made for data quality management will ensure continuous data quality improvement and management. These recommendations provide process improvement and corrective solution for data quality management. Process improvement measures ensures getting right data as early as possible e.g. use of standardized field lengths for first name, last name and other customer attributes across all systems. But if dirty data has been entered into the system the corrective measures will ensure fixing these data quality errors in a timely fashion e.g. implementation of SSN verification during Extract- Transform- Load process to identify invalid or duplicated SSN.

Chapter 9: Future Work

This problem needs enough work to be done in the future. Some of which can possibly be the following steps:

- Ontology based customer classification
- To study and compare the resources used in a neural network based classification vs rule based classification
- Providing services for the business layer: Develop rules to achieve strategic policies for business. Data mining tools can help to discover knowledge for such rules

References

1. http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
2. USHER: Improving Data Quality with Dynamic Forms by *Kuang Chen, Harr Chen, Neil Conway, Joseph M. Hellerstein, Tapan S. Parikh.* International Conference on Data Engineering - ICDE , pp. 321-332, 2010
3. Data Prediction in Manufacturing: an Improved Approach Using Least Squares Support Vector Machines. First International Workshop on Database Technology and Applications, 2009
4. Dataset provided by Nationwide Insurance
5. Anchoring data quality dimensions in ontological foundations by *Yair Wand and Richard Y. Wang. Communications of ACM, Volume 39 issue 1996*
6. An ontology-based approach to handling information quality in e-Science by *A. Preece, P. Missier, S. Embury, B. Jin, and M. Greenwood* . Concurrency and Computation Practice and Experience, May 2007
7. A Practical Guide to Enterprise Architecture *Chief information officer's council.*
8. Data Integration in Vector Databases by *Peter Buneman*
9. Handling Query Imprecision & Data Incompleteness in Autonomous Databases by *Subbarao Kambhampati, Garrett Wolf, Yi Chen, Hemal Khatri, Bhaumik*

Chokshi, Jianchun Fan, Ullas Nambiar. The Conference on Innovative Data Systems Research, 2007.

10. Application of Neural Networks for Customer Matching *Artika Agrawal, Jason Hursey and Chiu Ni Wang*

11. Principles of Lean Thinking *Mary Poppendieck*

12. QuickMig - Automatic Schema Matching for Data Migration Projects" by Christian Drumm, Matthias Schmitt, Hong-Hai Do and Erhard Rahm. Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, 2007

13. Best practices in data quality by *Salesforce*

14. www.salesforce.com/community/.../Best-Practice-Data-Quality.ppt

15. Steps for ensuring data quality: http://www2.ed.gov/about/offices/list/os/technology/plan/2004/site/docs_and_pdf/Data_Quality_Audits_from_ESP_Solutions_Group.pdf

16. Data management and data quality: http://www.iso.com/Research-and-Analyses/ISO-Review/Data-Management-and-Data-Quality-Best-Practices.html

17. Data quality best practices : http://www.dqguide.com/data-quality-tech-best-practices.html

18. Profit by data quality best practices: http://www.dataqualitypro.com/data-quality-home/profit-by-data-quality-best-practices.html

19. Data quality standards and best practices: http://www.datamentors.com/Figures/WhitePapers/data_quality_standards_and_best_practices.pdf

20. http://www.statsoft.com/textbook/neural-networks/

21. http://www.mathworks.com/

22. http://www.mathworks.com/products/neuralnet/

23. http://www.nationwide.com/

24. http://www.merriampark.com/ld.htm

25. http://en.wikipedia.org/wiki/Levenshtein_distance

26. http://wwwiti.cs.uni-magdeburg.de/iti_db/lehre/dw/paper/data_cleaning.pdf

27. Data Cleansing: Problems and current Aprroaches by *Erhard Rahm and Hong Hi Do*. Microsoft Research, Redmond, WA.

28. Using Data Mining Techniques to Discover Bias Patterns in Missing Data *Monica Chiarini Tremblay, Kaushik Dutta and Debra Vandermeer Florida International University.* Journal of Data and Information Quality, Volume 2 Issue 1, July 2010

29. Co- Engineering Applications and Adaptive Business Technologies in Practice by Jay Ramanathan and Rajiv Ramnath. Information Science Reference, 2009

30. Solving the data duplication problem for complex databases using neural networks, Florida Institute of Technology Melbourne, FL, USA 2003

31. Reasoning with Goal Models by Paolo Giorgini, John Mylopoulus, Eleonora Nicchiarelli and Roberto Sebastiani

32. http://www.opengroup.org/togaf/