Implementation of Lean Six Sigma® Principles: Making Data Cleansing Lean

Nityanand Wachche
B.E in Computer Engineering
University of Mumbai
Mumbai, India

ABSTRACT

Data cleansing is required before performing data load operation. It ensures that data loading on the system is organized, error-free and unique records. This paper puts the limelight on identifying issues and provides a solution for it, using Lean Six Sigma® principles. The solution presented helps to perform data cleansing operation accurately and cost effective for the project.

Keywords

Data Cleansing, Lean Six Sigma, Fishbone Analysis, DataLoad

1. INTRODUCTION

This paper is based on a business case where different formats of files are received from the business. For data cleansing operation requires 6 hours on the single data file and there are 5 files per month. This operation performed using Microsoft Excel® .To improve efficiency and performance, it is necessary to reduce the time consumed for the data cleansing operation.

2. DATA CLEANSING

Data cleansing operation must be performed before entering data into data collections because it helps to maintain accuracy and consistency in data collections. It deals with

- Detecting errors
- Removing errors and inconsistencies from data

The sources often contain duplicate data, invalid data format, and improper value or garbage value. In order to load accurate and consistent data, it is necessary to eliminate inconsistent information from different data representations [8].

3. LEAN

Lean is a systematic approach to identifying and eliminate waste (non-value-added activities) through continuous improvement in the process . The lean methodology was developed by Toyota Production System®. Lean mainly emphasis to increase the efficiency of the process and customer satisfaction. The core of lean is waste elimination [7].

Lean principles are not restricted to any particular field. The main intention of implementing lean is the improvement of the process. These principles are like management strategy. It can be implemented in various fields such as manufacturing, healthcare, and software development [2].

3.1 Lean Five Key Principle

Lean thinking distils the essence of the lean approach into five key principles [3]:

3.1.1 Value

Values are specifications of particular product or service which meet customer needs. Values are defined by the customer. These values should be fulfilled within specified time and price .Values specify the value in terms of service/product which is important from the customer's perspective .This principle depicts that lean is customer oriented methodology.

3.1.2 Value Stream

The Value stream includes complete process steps. Analysis of activities on the basis of values will differentiate activities into three categories which help to recognize non-value based activities .This principle will help to improve the process by eliminating waste.

3.1.3 Flow

After elimination of waste, it is essential to make a process step synchronized with value added activities. For creating smooth process steps, it is essential to make the flow of value added activities. This step helps to improve efficiency and hinder the creation of waste.

3.1.4 Pull

This principle states that value based process should execute only when customer demand, not prior to order. To fulfill the customer requirement on time great efficiency, the short cycle time of the process is a must. Create product only when customer demand is the principle of lean.

3.1.5 Perfection

Pursue continuous process of improvement to strive for perfection is the main attitude of lean. The main aim behind implementing lean is to achieve perfection in the process. This is possible by systematically and continuously removing waste.Continuous improvement process are moving towards perfection.

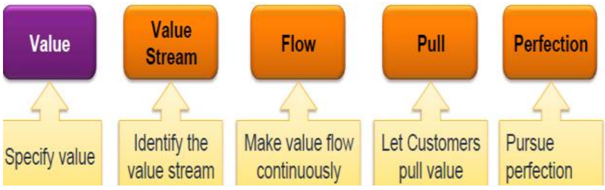

Fig 1: Five key principles of Lean

3.2 Types of Waste

When thinking about waste, it is useful to define the three different types of activity within your organization:

3.2.1 Value Added

Value added activities are those for which customer would be happy to pay for it.

3.2.2 Enabling Value Added

Enabling Value added activities includes quality checking, verification or inspecting product. Customers are not paying for this activity, but it is essential to maintain the quality of product delivered.

3.2.3 NonValue Added

Non value added activity includes that activity which is not contributing to improving the quality of product or service. These are unwanted activities which should be removed.

Table 1. Types of Waste

Types of Waste	**Example**
Waiting	Unassigned capacity between projects
Over Production	Development exceeding contract scope
Over Processing	Non-actionable supporting tools
Inventory	Large backlog of tickets
Motion	Resources switched between tasks
Rework	Code being revised
Transportation	Multiple handoffs in ticket management
Intellect	People at wrong place in the organization

3.3 Lean Methods

There are two primary pillars of the Lean system. The first famous pillar of the system is Jidoka (Build in Quality) and the second pillar of the system is Just In Time (JIT).

3.3.1 Jidoka

Jidoka refers simply to the ability of humans or machines to detect abnormal condition present in methods or process and to prevent the abnormality from being passed on to the next process. Jidoka highlights the identification of the problem and immediately stops the process when the problem is pointed out. The process flow is stopped until the issue is resolved. The main purpose of jidoka is to build the qualitative process rather than depending on enabling value added actions such as quality checking. Jidoka motivates to implement Automation [9].

3.3.2 Just In Time (JIT)

JIT aims for quick response with minimum waste and perfect quality. It focuses on improving process rather than spending the amount on storage goods. To implement JIT,it requires discipline and effective implementation.

3.3.3 Standardization

Standardization means creating a standard document, defining rules which act as a guide during process flow. This help to maintain consistency in the process with minimum variation. It ensures that continuous improvement in the process is moving towards perfection. Quality check list (which is used before deploying product customer) is one of the best examples of standardization. [4].

3.3.4 Pull System

The pull system eliminates over production by limiting production to those parts which are required for production. Lean opposes inventory system. Pull systems emphasized on Produce whenever there is demand.

4. IMPLEMENTATION BASED ON LEAN PRINCIPLE

4.1 Identify customer value

In order to go lean, understanding of what customers really value is rudimentary. To get project focused on customer need, defining a value stream inside the project is important.

Parameter which is customer values:

- Accuracy
- Efficiency
- Time consumption / File
- Cost

4.2 Analysis of Data Cleansing Steps

After a defining customer value that is desired, it's important to analyze the process. Analysis of current process helps to find out which steps hinder to fulfill customer requirement. The main aim of analysis to reach the root cause of nonvalue added activity and find out a solution by discussing with experts. A systematic approach towards analysis of process reveals multiple issues. The amount of manual intervention affects the amount of acceptable levels of data quality. Once a list of rules or standards gets builds, it'll be much easier to actually begin cleansing.

4.2.1 Visual-Lean Six Sigma: Analyze data cleansing steps

The complexity level of process analysis is vary according to the situation. There are many tools such as fishbone analysis which help to do complex analysis systematically and easily.

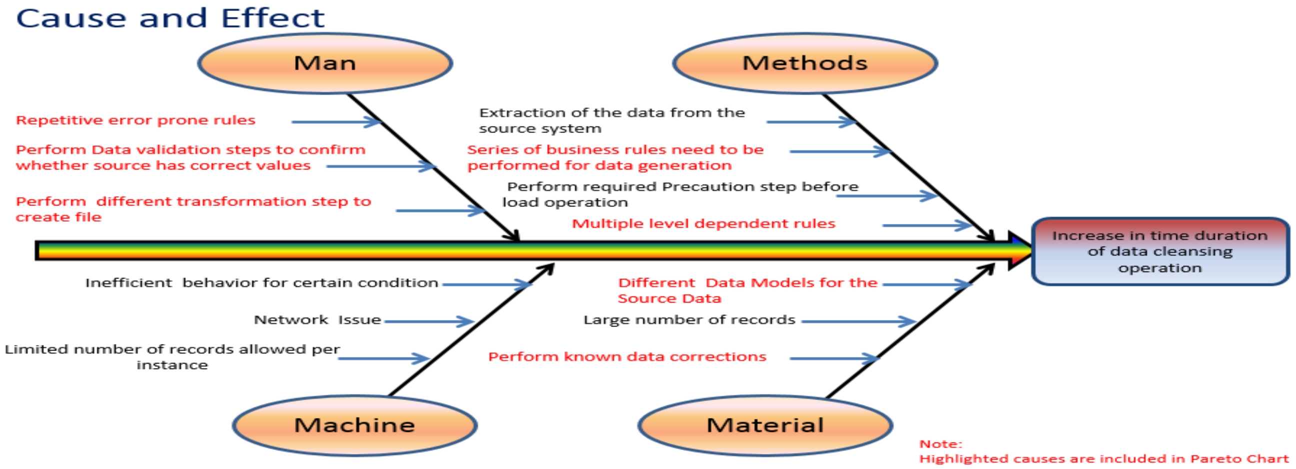
Cause and Effect
Man
Methods
Repetitive error prone rules
Perform Data validation steps to confirm whether source has correct values
Perform different transformation step to create file
Extraction of the data from the source system
Series of business rules need to be performed for data generation
Perform required Precaution step before load operation
Multiple level dependent rules
Increase in time duration of data cleansing operation
Inefficient behavior for certain condition
Network Issue
Limited number of records allowed per instance
Different Data Models for the Source Data
Large number of records
Perform known data corrections
Machine
Material
Note:
Highlighted causes are included in Pareto Chart

Fig 2: Fishbone Analysis of Case

4.2.1.1 Fishbone Analysis

The above Cause and effect diagram is known as Fishbone (Ishikawa) Analysis diagram.

The aim of Fishbone analysis must be to:

- Identifying causes of problem
- When team's thinking tends to dull and unproductive
- To analyze the issue in detailed level

Major categories of causes of problem:

- Machine (Equipment/Tool)
- People (Manpower)
- Material
- Methods
- Environment

The main branch of fishbone analysis shows primary categories for problem and branches show causes and sub-causes of the problem. The depth of diagram depends on the complexity of the problem.

Why does this happen? This question helps to find out causes and sub-causes of the problem.[5]

4.2.1.2 Pareto Chart

Pareto chart helps us to prioritize the issue, which highlights the areas on which we have to work so that it will create a major impact on the process. Pareto chart focuses on the most frequent or serious issue.

The Pareto chart provides a graphical representation of the Pareto principle, a theory states that 80% of the total problems incurred are caused by 20% of the problem-cause types

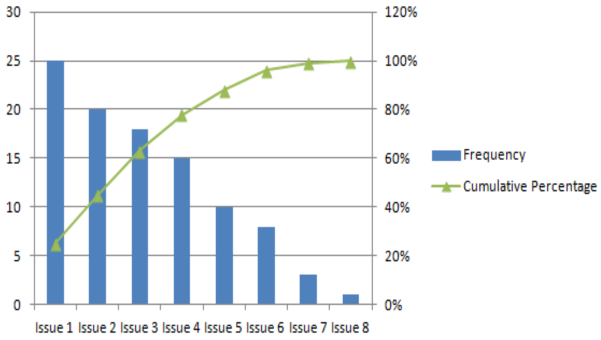

Table 2: Dataset of Parameter considered for Pareto Chart

Issue Number	Parameter	Frequency	Cumulative Amount	Cumulative Percentage
1	Repetitive error prone rules	25	25	25
2	Conditional Dependent rules	20	45	45
3	Perform Data validation steps	18	63	63
4	Series of business rules	15	78	78
5	different transformation step	10	88	88
6	Different Data Models for the Source Data	8	96	96
7	Tool Issue	3	99	99
8	Large number of Records	1	100	100

4.2.1.3 SIPOC Diagram

SIPOC diagram is a tool used by a team to identify all relevant elements of a process.

SIPOC Diagram include

- Suppliers: Who supplies required input (file) / raw material for the process
- Inputs: Key requirements are needed for process
- Process: Steps which are performed on inputs to obtain desired output
- Outputs: Result of process steps which are used by customers
- Customers: Recipients or users who use output generated by process

When to use SIPOC diagram:

- Working on project from scratch
- Defining Scope of Process Improvement
- For defining current situation before process Improvement

4.3 Implementing Automation

After analysis of current situation and factors which are obstructing to satisfy customer value, it is concluded that major efforts are consumed in performing data cleansing manual steps. Data cleansing operation is performed to meet some business rules. After analysis of the process,reason behind defect generation is performing manual steps to meet business rules. So the solution is to reach a goal is to standardize and automate data cleansing manual steps. These automated processes can run in real-time or in periodic intervals (daily, weekly, monthly) depending on how much data file.

4.4 Standardize template for Data file

Data standardization is a key part of ensuring data quality. Schema and integrity problems of data can be solved by declaring constraint or standard rule .There are a number of different ways to approach standardization, but it depends on which type of the data that's going into your system. The standardized rule ensures that properly formatted data should enter in the database. It helps to maintain consistency of data. [6].

Template of the data file which is received before data cleansing is required. Standardize template for data, is done by a collaborative discussion with the customer.

5. ANALYSIS OF RESULT

5.1 Analysis based on measurable parameter

Data cleansing operation is done before data load operation on the database. After analysis of current process shows that issue in data cleansing occurred due to

- Manual steps which were generates human errors
- Improper format or schema integrity problems

Analysis of result measured in effort time/month. Measure indicator represents how much time is spent on the individual data file. The result shows that effort time was reduced significantly after implementation of lean six sigma principle.

Table 3: Measurable Parameter considered for analysis

Measure Indicators	**Baseline Performance**	**Target Results**	**Achieved Result**
Effort time/Month	8 working Hours	6	6

Average number of files/month: 5

5.2 Cost Saving

Table 4: Report of cost saving

	Saving Details	
Pre - Improvement	Efforts per file (in hours)	8
	Number of files	5
	Time Consumed (in hours) / month	40
	Time Consumed (in hours) / year	480
Post-Improvement	Efforts per file (in hours)	6
	Number of files	5
	Time Consumed (in hours) / month	30
	Time Consumed (in hours) / year	360
Total Saving	Total time saved (in hours) / year	120

Inference: Amount of time required during data cleansing operation is reduced by 25%, which indirectly reduce cost as well

6. CHALLENGES OF DATA CLEANING

6.1 Limitation of Tool

Data cleansing process requires a tool to perform extract, transform, and load operation. Limitation of tool affects the efficiency of data cleansing process. If data collection contains a large number of records then tool may work slowly.

6.2 Identifying Process for Data Cleansing

Data cleansing is an iterative and explorative task. Data cleansing steps are interlinked and dependent.

The following five phases define a data cleansing process:

- Define and determine error types
- Search and identify error instances
- Analyze and find out solution to remove errors
- Correct the uncovered errors
- Perform preventive steps to avoid error in future

To find out the nature of errors and irregularity which are supposed to be eliminated, a descriptive analysis of data is necessary [1].

6.3 Preservation of Cleansed Data

Data cleansing operation is multiple step method. Preservation of output of previous data cleansing step is mandatory because it is a prerequisite for implementing next cleansing steps. Cleansing data is a time consuming and expensive task.

7. CONCLUSION

Data cleansing operation using traditional tools takes too much time and manual steps will affect the efficiency of the process. Improvement of the process is possible by effective utilization Lean Six Sigma principle. Lean six sigma® helps us to find non value added, waste, inefficient steps from the process. This methodology involves statistical and visual analysis to identify root areas of inefficient steps. Future scope of topic varies as per requirement. Furthermore, data cleansing automation can be performed using ETL tools like Talend, Abinito or SQL procedure.

8. REFERENCES

[1] Nidhi Choudhary, Department of Computer Science, UPTU, India, A Study over Problems and Approaches of data cleansing/cleaning

[2] http://www.institute.nhs.uk/building_capability/general/lean_thinking.html

[3] http://www.maskell.com/lean_accounting/subpages/lean_manufacturing/lt_the_principles_of_lean_manufacturing.html

[4] http://www.artoflean.com/files/Basic_TPS_Handbook_v1.pdf

[5] http://asq.org/learn-about-quality/cause-analysis-tools/overview/fishbone.html

[6] https://www.ringlead.com/blog/data-standardization/

[7] http://www.lean.org/WhatsLean

[8] Erhard Rahm, Hong Hai Do University of Leipzig, Germany http://dbs.uni-leipzig.de

[9] http://www.lean.org/lexicon/jidoka