LUISS

Department of Business and Management
Management and Computer Science

Course of DATABASES & BIG DATA

Impact of Data Quality on Business Intelligence: Strategies for Data Cleaning and Preparation

Prof.Blerina Sinaimeri
SUPERVISOR

Yllka Ostergllava 268121
CANDIDATE

Academic Year 2024/2025

Table Of Contents

INTRODUCTION

Deriving key insights from data is a fundamental requirement for the smooth functioning of corporations in the data age. Considering the potentially enormous amount of data that both RTALC and if GBT would produce on basis, enterprises mainly rely on Business Intelligence (BI) tools in order to analyze and understand this data and, thus, to make decisions. These tools were intended to make sense of data - to turn information into knowledge that is deployed and useful. Such knowledge can help in better management, better strategic decision and can give to a firm a competitive edge over other companies. The business utility of business intelligence (BI) is in proportion to the quality of the business data it is based on.

When there is no data to rely upon and data quality is weak, business intelligence data can paint the wrong picture and this corrupts decision making and leads to costly mistakes.
Quality of data is about the conditions of a data set, and is measured by factors consisting of accuracy, completeness, conformance, consistency, format, data area and data content, and other things according to a particular field of study. Good data is data that truly reflect the reality that it is supposed to represent and can be analyzed reliably. Conversely, poor data quality can be expressed in a number of ways – missing data, data that is out-of-date, data that differs in formatting in different sources, or data that is duplicated unnecessarily.The root of these problems can be: human mistake, bad system integration, and collecting data from lekty sources, that are poorly synchronized.

As data is the foundation for any analysis in business intelligence (BI), organisations need to put a great deal of focus on the data quality and accuracy, to be ready for analysis.
There are a number of potential issues from many sources, including human error, integration across systems, and combining data from multiple sources.
Because of the intrinsic significance of data in BI, organizations should focus on getting clean, reliable and analytically ready data. Data cleaning (or data cleansing) in business intelligence (BI) is of such importance that it is often viewed as constituting 80% of the work of building BI systems. This process is not just a technical exercise, but a critical element of the accuracy and appropriateness of the insights generated from the data. The early and most crucial phase in a business intelligence (BI) process is data preparation, which includes data cleansing, transformation, and integration. This feature is designed to make data preparation and data analysis more efficient.

Megalomania can spring from wrong or insufficient information causing damaging corporate strategies to emerge. For example, if for a company redundant or out-of-date data was entered into sales records, they may be analyzing the results and draw inaccurate conclusions regarding sales success that is not reflected in the current condition of the company. Similarly, fragmented data in an enterprise could lead to inconsistencies that influence decisions across individual departments. Therefore, reliable methods must be used for managing and cleaning the data.
An important issue in performing good data cleaning is the systematic approach used to discover and correct data quality problems.

This automated process may include using automated equipment to sense and adjust any errors as well as human operations (or the like) in which skilled operatives study results and make manual adjustments. The goal is to end up with a perfectly precise, neatly structured, and comprehensive data set that will provide a strong foundation for data processing using business intelligence software. Data preprocessing is the process of converting data into an appropriate form for analysis.

This may include combining data from different sources, normalizing data to make it uniform, and integrating new data to increase the accuracy and validity of the data. A top strategic asset is reliable data, which is a technological requirement for business intelligence. With a focus on data quality management, companies can ensure that their Business Intelligence (BI) systems are producing reliable business insights that reflect the reality of the organisation.

A data consumer is a role within this system and for the purpose of this article can be seen as a decision maker. Two other actors are the data producer and data custodian (Parker, Stofberg, de la Harpe, Wills and Venter 2006; Strong et al. 1997)

Consumer of data: A person who uses data
Data producer: Individuals/sources generating data
Data owner: People that have the overall responsibility for the data and supply the resources to handling (processing, storing) the data.

Chapter 1: Understanding Data Quality and Its Importance in Business Intelligence

An organization is foray into the fast paced Business intelligence (BI) world will either lead to substantial growth or costly mistakes, depending on the data it is using. Quality of Data in Business Intelligence Systems highlights the importance of the quality and effectiveness of the data in the business and management (operational or strategic) decision making processes within a company. This chapter is intended to bring to light the complex and dynamic nature of data quality in the context of business intelligence (BI), highlighting its role as a strategic asset that should be managed meticulously in order to realize its full benefits.

Standardizing data in business intelligence is a challenge where there is a need to maintain data accuracy, completeness, reliability, and relevance of data across its lifecycle. Each dimension has his own challenges and potentials for improvement of business intelligence, from the processes to the infrastructure. Thus, accuracy is important to ensure that the data accurately reflects what truly occurred, which is important for being able to accurately draw valid analytical conclusions. At the opposite, completeness guarantees taking all useful information into account so that the observation's perspective for analyzing is sufficient.

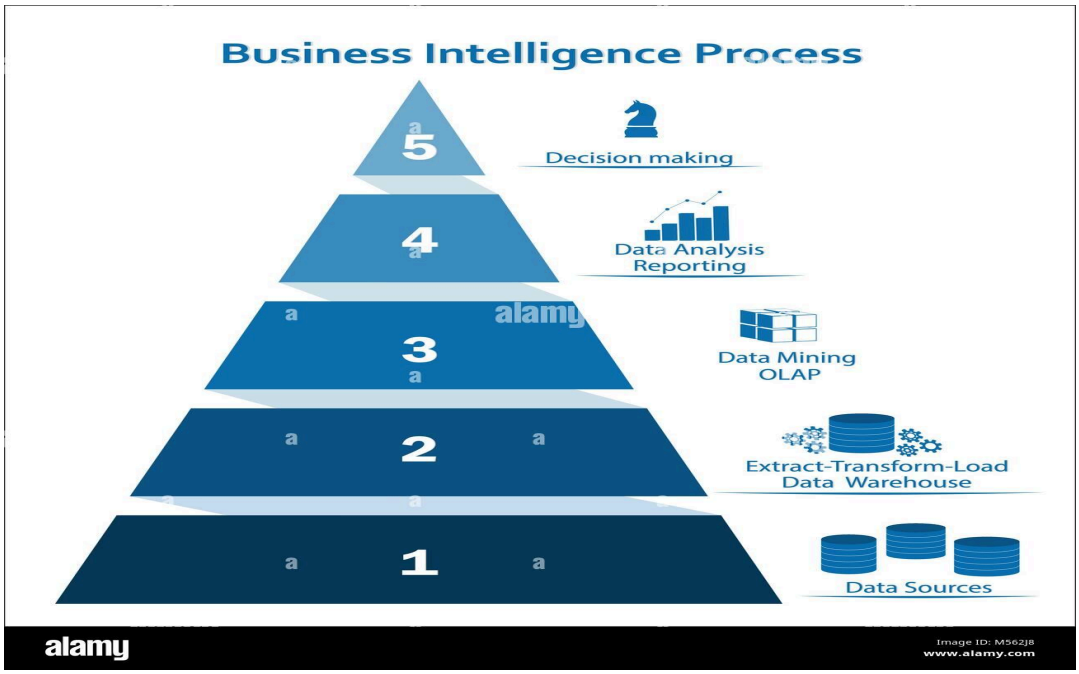

Figure 1.1: Core dimensions of data quality relevant to Business Intelligence systems.

The combination of heterogeneous data makes this more complex. It's common practice to consume data from many internal and external sources, each having their own standards and format. Such variation, aggregated in one business intelligence system, could lead to discrepancies, contradictory conclusions. The current study highlights the risk for data integrity, with potential gains that may lead to incorrect results, in cases where such integration problems are not adequately corrected.
To minimize these risks, effective data governance frameworks need to be established and implemented, which articulate and maintain rigorous data quality standards.

Also, the size and speed of the data continues to grow with the evolution of business intelligence (BI) tools.The robust protocols and systems able to be expanded and modified quickly are needed to process this large amount of data efficiently. This paper investigates the possibility of applying new technical approaches such as AI and Machine Learning models for the automatic detection of data corruption and the correction of such rot. What is more, these technologies not only reinforce the efficiency of data quality management methods, but also dramatically minimize the chance of human error.

Conclusion, the purpose of this chapter is to complete the considerations in the relevance of data quality in the context of business intelligence. Building on the foundation of "Managing Data Quality in Business Intelligence Applications," this two-part article focuses on specific technical solutions and strategic approaches to establishing superb data quality. The purpose is to show the importance of data quality as it relates to maintaining operational accuracy and increasing the organization's readiness to make data-driven, informed decisions that support the strategic goals of the business.

1. 1 Introduction to Data Quality

The quality of data is a characteristic of data management that includes many aspects and is also of significant influence on the business intelligence (BI) systems. There is a direct correlation between the data pumping into BI systems and the insight reliability, precision, and validity as well as the overall functioning of the strategies within the various organizational levels. In this section, we address the basic tenets of data quality and its impact on business intelligence as well as the early steps towards upholding basic data quality in BI activities.

Data Quality Evaluation

Quality is the degree of conformity of a data set to the demands set forth for the data set taking into account its functional use in process, decision making, and planning (Abu-AlSondos, 2023). This entails a number of aspects like accuracy, completeness, consistency, and relevance amongst others. Each facet plays a collective role towards the success of the data once it has been deployed for use. For example, the accuracy has to do with the fact that the data demonstrates the reality of what it represents, where as the completeness sets forth that no important data is omitted in order to make sure that the misleading aspects of the BI are prevented.

Importance in Business Intelligence

High data quality is crucial in the context of performing BI. BI tools make use of data to create analytic reports, which further help in the making of vital decisions in the operation of the business. If the data has a problem, the interpretations and the strategies derived from it are likely to be wrong and the organization is bound to incur some expenses trying to correct such things. For instance, if a retail firm depends on historical sales data that is missing certain trends, it might not know such a trend exists, and stock levels will not be set as required, resulting in sales being missed.

Challenges in Ensuring Data Quality

Data quality is becoming one of the toughest problems due to the sources of data and the volume of data that needs to be collected by the enterprises. Data can be derived from internal systems, social networks, external databases or even thousands of other origins. Each of these origins may have different levels of quality and data formatting.

Typical examples are:

a) Non-compatibility of information registered on the same data about other data or entry methods besides the data entered(Batini & Scannapieco, 2016)

b) Duplication of the entries of some data making it hard to interpret and overestimate(Christen, 2012).

c) Obsolescence of the relevance of the data captured making it necessary to periodically have them updated and validated (Pipino, Lee, & Wang, 2002).

Strategies for Data Cleansing and Data Preparation

It is of utmost importance to ensure data-cleansing and data preparation techniques and practices in order to increase data quality. Data cleansing is the process of finding and correcting any errors or inconsistencies in the data while data preparation is the process of converting the data into an evaluation-ready format.

Steps in this process typically include:

a) **Data Auditing:** Evaluating data to identify inconsistencies, redundancy, and other anomalies (Maydanchik, 2007).

b) **Rule Establishment:** The creation of rules on how the data in question should be treated and corrected(Hellerstein, 2008).

c) **Data cleansing:** Using software tools, as well as algorithms, in order to repair the anomalies found, for example deletion of duplicate figures or replacing missing information with estimates(Zhang et al., 2023).

d) **Data Validation:** Verification and testing of the processed data in order to confirm that the quality of the data is acceptable, and even to the better, standards as well as business requirements(Stonebraker et al., 2013).

e) **Technological Tools and Techniques:** There are a variety of software tools and other technologies that are used to address data quality problems. This consists of a range of data profiling tools which examine data quality, data cleansing applications which function mechanically to resolve data quality issues, data quality procedures which are valid in ongoing data quality activities. A trend towards the usage of technology two types of technology has emerged, which is aimed to both replace and heighten the effectiveness of data cleansing within the processes(Microsoft Research, 2021).

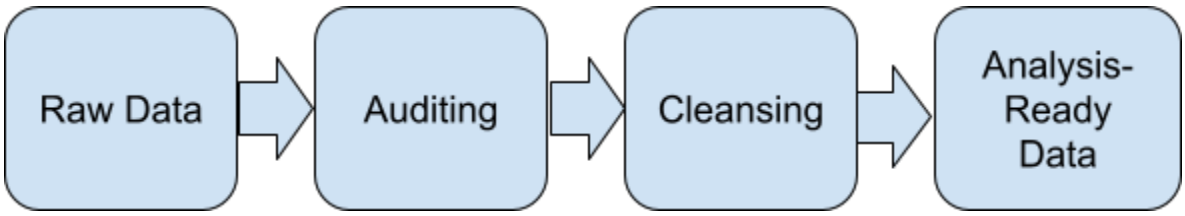

Figure 1.2: Data quality processing flow from raw input to analysis-ready output.

It is the quality of data in business intelligence systems that largely influences the quality and usefulness of interpretations derived from them. As more companies opt for data-oriented approaches in order to stay competitive, the value of proper cleansing and preparation of data increases in BI practices. Ensuring good quality of data is not only an operational requirement but a vital component that supports the decision making in the world of business analytics.

1.2 Data Quality and Business Intelligence

It is needless to explicate that the latent correlation that exists between data quality and business intelligence (BI) is profound. It is data quality that forms the backbone of information systems such as BI, where insights and decisions that are reached are correct and have actionable purposes. In this part, we will study the relations between data quality and BI providing examples on how data quality plays an important role in enhancing the activities of the business and making informed choices.

Impact of Data Quality on BI Systems

The degree of efficiency and effectiveness of the BI systems is a function of the level of data quality possessed by the systems. High data quality gives rise to high data analytics which in the end gives the stakeholders reason to have confidence in the decisions made. Poor data quality however can be detrimental to the decision-makers since it can cause poor decision making and implementation which will harm the overall business. For example, the management data may be so poor as to persuade

management into committing valuable resources to develop a product which the market will not accept. (ICIQ 2016, Article 4, Publication date: June 22nd, 2016)

Item	Related Studies
Data consistency ("Single Point of Truth")	Bachmann and Kemper [2011], Turban [2011], Morabito et al. [2011], Lang [2015], Hartl et al. [2016]
Data integrity during simultaneous use	Negash [2004], Bachmann and Kemper [2011], Lien Mbep and Jacob [2014], Hartl et al. [2016]
Traceability of BI master data changes	Seufert and Oehler [2011], Lang [2015], McKnight [2014], Hartl et al. [2016]
24/7 operation of the BI system	Turban [2011], Morabito et al. [2011], Andriole [2006], Hartl et al. [2016]
BI architecture is documented	Wieder et al. [2012], Kemper et al. [2010], Lien Mbep and Jacob [2014], Hartl et al. [2016]
Only mandatory BI tools are used	Akhavan and Salehi [2013], Lien Mbep and Jacob [2014], accenture [2007], Hartl et al. [2016]
Responsibilities for the BI-development are clearly shared between the company's departments and the IT and communicated throughout the whole enterprise	Bachmann and Kemper [2011], Hartl et al. [2016]
User rights to access the BI system are defined	Eckerson [2008], Kemper et al. [2010], Lien Mbep and Jacob [2014], Hartl et al. [2016]

Table 1.1 Data quality and provisioning items that impact BI

Data Quality Dimensions Affecting BI:

a) **Accuracy:** This resource highlights that accurate analysis, which in turn leads to reliable business decisions, can only be powered by high-quality data. Higher processing costs and unreliable analysis can result from poor data quality. (Collibra, n.d.; DATAVERSITY, n.d.)

b) **Completeness:** Missing values may result in deals that cannot be fully explained and certain factors may be ignored. For instance, if the voices of the customers do not reach the management efficiently there are chances that the company remains the way it is and refrain from further enhancement of its products.(Pipino, Lee & Wang, 2002)

c) **Consistency:** the same names and formats are present in every dataset. Inconsistencies could cause confusion in implementation order or the measurement of performance, thereby influencing the evaluation of IT ventures (Pipino, Lee, & Wang, 2002). For instance, if the same KPI is referred to by two departments with different names, the dashboard can end up having conflicting data, further eroding trust in the analysis.

d) **Timeliness:** In cases where information is out-of-date it is more likely that irrelevant decisions will be made or tactical planning and operations cannot be responsive. (Precisely, 2023)

Operational Efficiency : Adequate data enables organizations to enhance their processes and cut costs as well as enhance service levels through effective forecasting and optimal utilization of resources.Firms that use big data achieve over 10% cost reductions in operations (McAfee & Brynjolfsson, 2012).

Customer Satisfaction
Therefore, the business will utilize this information to understand customers' needs better and serve them in a way that will not only attract them to make purchases but also retain them.

Regulatory Compliance
Several industries require comprehensive management of data that signifies the need for compliance with adequate data quality to prevent penalties associated with defendant reputations.

Competitive Advantage
However, high data quality helps the companies outperform their competitors by pinpointing market trends, customer needs, and ways to improve business operations faster.

Challenges to Maintaining Data Quality in BI
Same Day Paper Two other specific problems relating to the information lifecycle involve poor management of vast sets of business information and poor resiliency of quality information over time. Data, however, is dynamic and can change with time. When organizations grow and become more complex, it becomes a challenge to keep data consistent and accurate when there are more data sources to operate with. One such difficulty is fusing data from many disjointed and redundant repositories into one coherent BI system without losing data quality, and effective control mechanisms are well integrated in these systems.
Best Practices for Ensuring Data Quality in BI: About fedex Data Quality Reporting Or Polling Threshold: Also are DQA safety nets. Only such approaches which will recommend standardisation and unification of DQ reporting processes.

Security Frameworks
Security policies concerning the physical and logical safeguards to information resources such as data sources, databases and applications are documented and generated also.

Regular Audits and Updates
Another approach to workplace intelligence is regular and constant quality audits maintained by employees since there must be enough emphasis and training on accurate or relevant data entry to avoid data obsolescence.

Use of Advanced Tools
Act for Anti-corruption Policy Policies that deal with internal procurement processes and safeguarding procurement resources are fundamental while seeking to ensure that data is of high quality.

Conclusion, as such, there is a strong relationship between the quality of BI information and social electronic business performance against traditional businesses. Therefore, any business intelligence systems should not be put in place without ensuring that the data is of good quality. It is crucial for each organization to deal with the quality of information if they want to make the best out of their business intelligence systems and hence proper and effective operational and strategic decisions are made based on the insights provided by the BI systems.

1. 3 Summary of Chapter Goals

Scope for Understanding the Quality of Data
Business intelligence, like any other tool within an organization, is hampered by the quality of the data and therefore the information derived from that tool. A strong data quality comprises the attributes of the data being right, sufficient, current, and useful. Rightness is the data portrayal of the real constructs it signifies. Sufficiency is the absence of gaps that could induce wrong actions. Currency ensures the information is current in correspondence with the most recent data while usefulness ensures that the data is appropriate to the issues being investigated by the organization/s.

The framework assisted in understanding the relative dimensions of these concerns by exploring each of their inter-related features as well as the respective ensemble effect on the performance of the Business Intelligence Systems. For instance, it is crucial to ensure accuracy and completeness of patients' records in the healthcare industry, to enable appropriate treatment and subsequent billing. Such errors or omissions could result in costly medical or financial mistakes.

Data Quality and Business Intelligence
In this section, the direct relationship between the data quality and Business Intelligence (BI) systems is explored. Appropriate data leads to valid analysis which streamlines business processes and facilitates sound decisions. That is the reason why when the quality of the data is low, its utilizers will hold wrong information and make decisions which may lead to a lot of loss in terms of money and reputation.

A good example presenting the dependency between data quality and BI is adoption of customer purchase data in a retail company for determination of optimal inventory levels. The analysis relies on accurate data (which may be input errors made at point of sale) to ensure that the company orders are not huge stock or limited.

Integration Challenges and Solutions
At the same time, the issue of ensuring quality is made more difficult by integration of data coming from various sources. Every single source might establish different standards and formats which may lead to errors and variance where all of them are incorporated into a single BI system. In this section, data integration challenges are evaluated and the key role that data governance structures play to uphold integrity and consistency in datasets are explained.

A real life example therefore would be a global corporation which extracts sales from different countries and consolidates the sales data. Foreign exchange issues, legislation differences on taxation, and different methods of recording sales can all lead to differences in the global sales reports unless they are faced.

Advanced Technological Interventions
The chapter also discusses why introducing other advanced devices such as the artificial intelligence (AI) or the machine learning (ML) tools that can be used to improve and maintain data quality is warranted. For example, these technologies have the ability to perform data error detection and rectification automatically. This capability minimizes the occurrence of human errors and maximizes the effectiveness of the data quality control processes.
For instance, AI-based transaction monitoring systems and techniques can include businesses' use of AI software that systematically detects anomalies in transaction information. As a result of these aberrations being corrected mechanically, a business is able to enhance its accuracy in financial communication and compliance with government legislation.

Objectives of the Assessment: Underlining the importance of Data Quality in business intelligence is the primary motive of this chapter. The consensus is reached on the importance of data quality management in the entire process such that it would create knowledge of high impact for the purpose of ratings: how high-quality data is a must-have requirement for a productive BI system. This understanding is key in organizations that seek to make data based decisions as they formulate strategies and wish to sustain their level within the business environment.

To conclude, in the last section of the chapter it becomes clear that the doubts of maintaining such high standards of data quality are numerous but their advantages are even more. The present study helps to further develop the theory related to the quality of information. The management of data quality measures improves action and efficiency risks, client satisfaction, and regulating requirements of performance and therefore enhances the general firm health.

Chapter 2: Challenges and Current Practices in Data Cleaning

Challenges and Current Solutions to Data Cleaning
As data volumes grow and more complex and diverse data is collected by organizations, data cleaning, which is the process of enhancing the quality of data, entails several challenges to organizations. The accuracy, completeness, and consistency of data validation are some of the main aspects that determine the functionality of business intelligence systems. It is noteworthy that many organizations have some commonality being data duplication, lacking data, and varying formats. Such problems are often exacerbated by the potential and reality of combining multiple data sets which require cleaning after processing because of the inconsistencies created.

This chapter assesses the pivotal challenges that organizations face in data cleaning and summarizes the relevant methodologies that have been proposed to address these challenges. While some enterprises are still using human data cleaning processes, gradually others are adopting and implementing technological solutions that make use of machine learning and artificial intelligence for data quality issues. The aim of this chapter is to increase understanding of the role of data cleaning in the decision making and business intelligence processes of the organization through analysis of the limitations and effectiveness of the methods currently in use.

2.1 Overview of Data Cleaning Challenges

Data cleansing is a vital activity for all data used in analysis and decision making. However, data cleaning has its drawbacks, and, if neglected, it contributes to poor quality datasets. Some of the basic challenges are summarized below.

Missing Data
Online data cleaning issues scrutiny Data scrubbing is an essential step in any dataset used for analysis and decisions. Yet, data tidying has its limitations, and if overlooked, it leads to low quality datasets. Some of the main obstacles are reviewed below. Missing Data One of the most frequent issue you will encounter during your data cleaning battle is 'presence' or 'absence' of certain fields/values in your dataset. This may be attributed to various causes including errors in data entry or lack of completeness in procedures. Incomplete data might lead to bias in the analysis, reduced prediction model score and reduced amount of usable data. Ignoring missing values usually yields the incorrect results that can not be trusted.

Inconsistent Data

In the context of data cleaning, data inconsistency is where there are two or more representations of the same information in the database. Examples include variations in dates: some dates may be written in a MM/DD/YYYY format while others may prefer DD/MM/YYYY formats; sometimes even the language employed is different or measurement units do not match, and so on. These challenges affect the overall data consolidation and analysis because incorrect assumptions may arise due to conflicting data or lack of a well-adjusted data comparison analysis. The solutions available include ignoring these problems and the wearing of standardized data formats.

Duplicate Data

Duplicate data refers to the same record being found more than one time in any single dataset. This usually happens through data entry mistakes or problems with integrating two systems. This redundancy not only increases the amount of work but it also affects the statistics and their usage in predicting the outcomes and may alter the comparisons-positively or negatively. Duplicate record identification and removal is one of the many ways to maintain the reliability and accuracy of the information collected.

Erroneous Data

Erroneous data is that information in the databases which is misleading or wrong. This might be due to many factors such as wrong database management, mistakes in findings from records or bad input of data in the databases. This challenge is very risky since worthless information will damage the essence of the analyses obtained and therefore lead to the wrong decision making. Examples, shows there can be a wrong entry of customer address fields, value on transaction fields like amounts, date & so forth. It is therefore necessary to have mechanisms such as validation of rules or automated error checker tools which help to minimize the wrong data.

Outliers

Outliers are considered anomalous as they deviate substantially from other observations in the sample. Although outliers may include actual observations in certain cases, these are more often than not the complications of the research study or some aspects of its data. If not properly managed, outliers will affect the integrity of statistical models, the validity and reliability of analyses, and even the forecast of outcomes. Some solutions therefore must include outlier detection and treatment techniques to alleviates the effect of these values on outcome analysis.

Data Integration Issues

Data from multiple sources is integrated with great difficulties as sources may adopt different data structures, formats and even naming conventions. These discrepancies, when pieces of information are joined from various data sets, could bring out and mismatches of data, data inconsistencies and lack of data compatibility which then would compromise the accuracy of the analysis. Solving such data integration problems will need some transformation methods, data reconciliation, and also harmonization methods on the integrated dataset.

Large Data Volumes.

The exponentially large size of the datasets makes data cleaning extremely hard. In cases of enormous datasets, efficient processing of such data calls for adequate amounts of computing power, and traditional methods of data cleansing cease to be effective as data volume increases. It is also important since when large quantities of data are processed, the chances of committing a ‘mistake' or encountering ‘inconsistencies’ also increase. Development in technologies has led to the formulation of data cleaning strategies that will cope with the current and the future demands of large datasets.

Noisy Data.

The noisy data is an inclusion of any information that is irrelevant to the activity being performed and/or is useless (whereby useful means relevant towards the goals) in the dataset providing useful information and revealing certain patterns. As usual there is a catch, such situation is especially visible in unstructured data, as in case of text data, there are invariably situations of further typos, needless symbols or filler content that reduces the quality of such data. Therefore, getting rid of noise and extraneous information is crucial for the correctness of these models and for successful analysis of the data collected.

Data Timeliness

Some applications require that the data collection process and the data cleansing processes occur in a reasonable amount of time. A delay in data cleansing leads to such data becoming old and losing its utility in prompt actions and analyses. For instance in the real time fraud detection systems, data needs to be cleaned almost in real time if proper fraud detection and prevention has to be attained. It is often necessary to implement automated data cleansing pipelines in order to meet these strict timeliness requirements.

Privacy and Security Concerns

The data that is cleaned is often sensitive, roofs for example, are cleaned to protect or include sensitive information, financial information, or healthcare information. Protecting such data and people is a high risk to undertake as there is a chance of wrong handling leading to shedding of sensitive data or breaking the law, for instance, the GDPR law. Data protection is enhanced through the use of data anonymization, data encryption and other invitatives that require complying with laws.

2.2 Review of Current Data Cleaning Practices

The term ‘data cleaning’ or ‘data cleansing’ encompasses a process that critically evaluates the correction of different datasets in terms of their accuracy, coherence, and quality in relation to data analysis, machine learning, as well as business intelligence. As industries become more and more hinged on making decisions on the basis of data, data-cleaning skills have tried to keep up and aim to fix different issues such as lack of values or no data at all, inconsistency, duplicates, and incorrect data. In this particular review, the author highlights the major areas and ways in which contemporary data cleaning is carried out.

1. **Handling Missing Data**

The problem of missing data is considered to be one of the most common problems faced by the researcher when cleaning the data set.Missing data is a common issue in research, accounting for up to 80% of datasets in some domains (Schafer & Graham, 2002). Practices aimed at managing missing data typically include the following approaches:

- **Imputation:** This refers to the practice of substituting a missing item by the statistical measures mean, median or mode of other missing items in the dataset(Little & Rubin, 2019; Van Buuren, 2018).
- **Deletion:** In some instances, when the extent of missing data is minimal, then those records with missing values can be excluded completely(Allison, 2001).

2. **Standardization and Normalization Strategies**

When dates are recorded using inconsistent values (date format, units or levels), there is the possibility of inconsistent data. Practices to address such discrepancies include:

- **Standardization:** Modification of data so that useful information conforms to a single convention. For instance, all date formats raw data can be changed into one specific date format e.g. all in DD/MM/YYYY and so forth (Han et al., 2011, Ch. 3).

- **Normalization:** This, typically, is the practice of transforming the values of a numerical feature into proportions, particularly helpful where algorithms which work on the data become sensitive to the volume of the difference between features.

3. Removing Repetitive Records

Information in a dataset may include duplicate records that come from different data sources or possibly by human factors such as typing errors. For this reason and for the purpose of enhancing precision and efficiency reducing the duplication of data is emphasized on the data cleaning:

- **Deduplication Algorithms:** Automatic systems are also available to remove duplicate records based on specific definitions and cutoffs. The records would be assumed to be duplicates, where the names or addresses do not necessarily match, eg in fuzzy matching (Christen, 2012).
- **Manual Review:** In some situations, however, deduplication can be more involved and may include looking at documents by physical means as deduction acknowledgement algorithms may not be definitive(Baxter et al., 2003).

4. Regarding Wrong Data

Errors that originate from manual input mistakes or any system error can lead to difficulties with regards to the quality of the dataset as erroneous data exists. In order to minimize the problem, the following practices are mentioned for use:

- **Validation Rules:** The practice of imposing restrictions on the marital status of an individual without discretion after contracting an African marriage would be regarded as an error. These rules can be checked at the data input stage and hence avoid the chances of collating such details in the system in the first case. For example, entering a date in a particular dd: mm: yyyy format or entering a numeric value within a certain range can prevent the wrong information from being entered to the system(Rahm & Do, 2000).
- **Error Detection Algorithms:** Systematic analysis can be performed to check for variance and deviation from typical values and practices as per predicated norms. For instance, extreme numbers or unreasonable numbers such as the age field that can have negative integers are put on alert for repairs to be conducted(Chandola et al., 2009).

5. Dealing with Outliers: Detection and Management

Any value that is poles apart from the sample can lead to distortion of some forms of analysis and hence introduces errors in data analysis. The present practices provide a number of techniques in dealing with outliers:

- **Statistical Methods:** Such thresholds for the discovery of outliers involve Z-scores (Aggarwal, 2017), the Interquartile Range (IQR) or their combination relations such as IQR-Z Scores.
- **Transformation or Removal:** In certain situations, such outliers need to be changed (transformed such as log transformation to limit their dominate influenced powerful values) or even eliminated if they are found to be wrong or irrelevant.

6. **Data Integration and Harmonization**

As organizations begin to appreciate the significance of big data, it has also necessitated the combining of data from several sources such as databases, worksheets or even cloud storage locations. Among important techniques in combining data without losing consistency are:

- **Data Mapping:** Under this practice, the various data elements coming from multiple data sources are cross-referenced. Note that when two fields correspond with each others, this means that the "mapping" is executed properly. Data mapping is used so primarily when data from different systems has to be combined(Lenzerini, 2002).
- **ETL Processes:** The Extract, Transform, Load or ETL method (Vassiliadis, 2009) is a standardized methodology that most people utilize to integrate and cleanse scattered large databases. In this step, known as the transformation step, the information is scrubbed and purified so that all data conforms to a standard in preparation for upload into a database.

7. **Automated Data Cleaning Tools**

With regards to size and complexity, the expansion is accompanied by growth in datatypes which would render data cleaning manually with estimates unrealistic. Data cleaning is mostly performed with the help of various technologies nowadays. Those include, but are not limited to:

- **OpenRefine**: An application that aids in the cleansing and converting of datasets. Ideal in dealing with large amounts of dataset entities where there a presence of errors and repeated records(Huynh et al., 2013).
- **Trifacta:** A software for performing chores as data cleansing and preparing data for analysis and providing graphics in the proper context to help the users in the data cleaning process(Kandel et al., 2011).

Libraries Supporting R and Python: In programming, for example, Python and R, libraries such as pandas and dplyr are available for users to write personalised cleaning codes for the datasets in use.

8. **Data Protection and Information Security Approaches**

Considering the current trend where there are a number of data protection laws that have been enacted such as the GDPR, measures taken during data cleaning must also address the aspect of privacy and security. Key practices include:

- **Anonymization:** During the cleaning procedure, many sensitive personal identifying information such as names, social security numbers, email addresses, and the like are anonymized so as to meet the purposes of the law (El Emam & Arbuckle, 2013; Sweeney, 2002).
- **Data Masking:** During the cleaning exercise confidential information is either masked, or encoded so as not to allow for leakage of sensitive personal information.

9. **Data cleaning on the fly in real time**

In systems that are real-time like financial markets, data cleaning has to be instant. The real-time developments in data cleaning include the following practices.

In this case, once the data enters the system, cleaning takes place simultaneously with the flow of the data through the system this is accomplished using the To this end make use of the Apache Kafka or Apache Flink editing which guarantees that consistency and validation of the data is touched on before any downstream processes.Technologies have been developed in real time whereby cleaning and verification of data is done with minimum delays, which eliminates processed data and the time lapse between collection of data and analysis.

2.3 Case Studies

Case Study 1: Retail Company – Handling Missing Values

Around the world, one of the largest retail companies with a customer base database faced the problem of missing fields in the picture of sales. Incomplete data sets were present in customer demographic and purchasing history and transaction time stamps.

Problems:

- **Incomplete Customer Profile:** The strange demographics and no demographic fill-in, meant that the customer profile was not complete hence limiting avenues for marketing.
- **Disjointed Sales Report:** Lack of transaction data or gaps in the transaction data led to faulty sales conclusions and forecasts which in turn interfered with stock control and financial plans.
- **Completing Deficient and Effectual Customer-care:** Apart from incomplete facts, even in the available information, details for specific needs of customer support could not be assured.

Practices Implemented:

While the company utilized other qualitative techniques of imputation like triangulation, for the quantitative uttered alternative mean substitution for numbers and most common mode for the respective variables were utilized.Merging external demographic data with internal systems necessitates thorough validation to solve quality and compatibility concerns (Loshin, 2013).

Data Completion: Organisations that employ rigorous audit standards reduce data errors by 40% when compared to ad hoc approaches (Redman, 2016).

Results: The retail company reported an enhancement in the precision of its customers' profiles and sales predictions. The effectiveness of targeted marketing improved and quality of customer service was enhanced as there were more detailed records.

Case Study 2: Healthcare Provider- Solving Data Conflicts

Background: The healthcare providers had issues on the standardization of data across their electronic health record (EHR) systems. In this case, this includes differences in the way date formats are recorded or the medical terms used.

Challenges:

- **Integration Issues:** The variance in data formats and terminologies made the incorporation of both EHRs and e-clinicals from various departments and systems arduous.
- **Clinical Decision support:** Data inconsistencies were detrimental to clinical decision support systems as accurate and consistent data is required.
- **Regulatory Compliance:** There were regulatory standards and reporting requirements that were set which required adherence to any record by accurate and consistent data entry.

Practices Implemented:

- **Standardization:** The provider embraced some data standardization practices, one of them is use of similar date formats to avoid loss of information (Batini & Scannapieco, 2016).
- **Data Transformation:** Such data transformation processes were introduced for the purpose of converting and standardizing data from different sets prior to integration (Kimball & Ross, 2013).
- **Continuous Monitoring:** There were subsequent processes of monitoring for those impacts and validating them to detect and address any inconsistencies (IBM, 2021).

Why were these system enhancements necessary?Integration of data by the healthcare provider led to improvement in quality of clinical decisions made and reports produced. Standardized data upholds the standards and regulations put in place ensuring better management of the patients.

Case Study 3: E-Commerce Platform – Handling Duplicated Customers

As customer growth of an e-commerce platform progressed impressively, problems of duplicate customers' information arose. Situations where there were two or more records provided for the same customer created problems in processing orders, segmenting customers, and suggesting products.

Challenges:

- **Order Fulfillment Issues:** Many places where customers were having duplicate records created multiple-order problems during order fulfillment and refused order inventory records.Retailers lose $20 for each duplicate record because they have to pay for shipping and restocking (Experian, 2022).
- **Segmentation Errors:** Companies' marketing activities were hampered due to wrong customer targeting arising from duplication of customers' data (Kumar et al., 2019)..
- **Customer Experience:** Customers spoke of problems related to having multiple records for the same customer which led to a poor customer experience (Accenture, 2020).

Practices Implemented

Duplicate Detection Algorithms - The platform used run of the mill sophisticated duplicate detection algorithms that employed fuzzy matching techniques for the identification of duplicate records and later merging them.

Data Cleansing Routines - Upon achieving the intended objectives, additional regular data cleansing routines were introduced to the system so as to discover and deal with duplicate records on a continuous basis.

Customer Verification - During registration, more measures regarding customer verification were introduced to prevent the creation of redundant records.

Results – The e-commerce platform achieved higher order accuracy, more effective marketing etc. While reconciling all the information into a consistent picture, the number of duplicated records decreased. This subsequently helped to optimize processes and increase customer satisfaction.

2.4 Summary of Challenges and Practices in Data Cleaning

Missing Data: Any missing data can hinder the effectiveness of data analyses and the validity of the findings. None or limited data can be caused by bad data or even procedures like data entry, system or software flaws or even by lapse in the data gathering exercise. The main question to be addressed is ‘what is the best strategy to adopt when data are absent without affecting the existing insights and biases of the analysis?’

Data Inconsistencies: These discrepancies usually pose a certain level of difficulties and obstacles in the integration of data that would otherwise be useful for analysis and interpretation. These discrepancies are also reasonably common in systems with several data sources or separate systems that have adopted diverse conventions. People always assume that all the datasets will give the correct representation required, without taking the necessary steps to ensure this is the case.

Duplicate Records: When data entry is monotonous for long periods of times, it can lead to data inaccuracies and eventually duplication. Duplicates are more often caused by system glitches, too many entry points for data or even excessive inconsistency in keeping of records. Keeping track of and consolidating duplicate data records is important for data quality.

Data Quality Issues: Lack of specific data on the market issues at that particular time, wrong or too old information, excessive data among others contribute low quality decisions derived from data. It is necessary to ensure data accuracy, currency, and relevance which calls for well-structured data accuracy and data cleansing processes.

Scalability: Increasing the size or complexity of datasets also increases the difficulty of cleaning. With the amount and volume of data aggregation high, performing the data cleaning tasks become more and more challenging. This challenge warrants large-scale approaches and even automation so as to manage large data efficiently.

Summary of Practices: Data Imputation: For missing values in a dataset, techniques from very simple mean, median, or mode replacing to more advanced ones like regression and machine learning based imputation methods are the ones done. These practices are meant for filling the holes in such a way that there is minimal change to the dataset’s structure and its intended usability.

Data Standardization: This step helps reduce the variability by the same set of data attributes by defining the data formats and terms. This practice involves the use of standard formats for dates, units, categories, controlled vocabulary or ontologies to databases to avoid terminological discrepancies.

Deduplication: Most organizations rely on algorithms and tools to help them identify duplicate records and resolve such problems. Duplicate records are resolved and synthesized using fuzzy matching, rule based deduplication and manual processes, thus enhancing accuracy and reducing redundancy.

Data Cleansing Routines: Continuous data cleansing routines are developed to the management of data quality over periods of time. Such routines embark on frequent assessment, validation, and rectification of data in order to correct errors, inconsistencies and outdated information. It is recommended that certain automated tools and scripts be used for streamlining these processes.

Data Enrichment: Activities which seek to incomplete or inaccurate records by renowned data enrichment involve the insertion of additional information from other external data sources to the existing one. This activity is useful because it assures that most details that are required for analytics are present thus facilitates better analytics.

Scalable Solutions: Solving issues of scalability requires the deployment of efficient data processing technologies and advanced automation tools. Such solutions are mostly cloud based as well as employing big data technologies that easily enable the processing of volumes of data in a timely manner thus allowing efficient data cleaning processes to be undertaken on a grand scale.

Conclusion: Data cleaning is critical in providing confidence in the data and the decision-making processes that will be performed on such data. Inconsistencies, dead volume, duplications, and data quality becomes a challenge that does necessitate exclusivity in terms of practices and approaches. To meet these goals, organizations need to apply relevant data imputation, standardization, duplicate detection, cleansing routines, enrichment and scalable solutions.

Chapter 3: Developing Effective Data Cleaning Strategies for Business Intelligence

Readers will understand why it must be emphasized that there is no business intelligence without data integrity. This chapter examines how information strategies pertaining to the quality and integrity of data can be incorporated into the process of cleaning datasets. The focus is to specify the particular guidelines that require consideration if the clients are to be provided with adequate data cleaning processes that do not only solve the data problems but are consistent with the organization's strategic objectives.

A proactive and responsive strategy to changing business and customer environments is adept at this. Data cleansing is such that it is regarded as a multi dimensional problem that needs the right combination of technology, process management, and business intelligence gathering. It is a field that encompasses the careful detection, adjustment and elimination of mistakes and discrepancies from data which makes the end product for business intelligence useful and purposeful. These strategies would be related to the understanding of the whole range of the data lifecycle within the analytical process.

The first step in developing these strategies is conducting an in-depth assessment of the existing data system within the business organization. This involves determining the sources of data acquisition, tracking the movement of data through the different processes in the organization, and highlighting the critical areas where data quality is always compromised further to understand their contributing factors. Such analysis will point to the relevant areas that need such focus by way of issuing inflammatory procedures on those volumes for instance such as removing duplicates, correcting mistakes as well as addressing the absence of certain values.

On the completion of the analysis of the critical areas of concern, the second stage in the data scrubbing process deals with the determination of the proper methods or solutions to be utilized in the cleaning exercise. The exceptionally wide range of the solutions that have been strides, would include, data scrubbing tools but also include very advanced AI systems that are able to predict when risk for data quality maintenance is low. Policy formulation on the appropriate technology should be informed by issues like the extent of data, levels of complexity of data sets as well as the actual types of errors to be fixed.In addition, the human factor must also be taken into account with regards to the undertaking of data cleaning strategy development. It is important to provide training and engage the people who use the data: in IT, in analytics or in operations. They have to appreciate the necessity of data quality and be trained to operate the tools selected for the purpose. Such an approach guarantees that data cleansing is not performed occasionally, rather, it becomes part of the culture and values of the organization.

Finally, these strategies ought to be put into action and then, as most things, must be continuously improved. This means introducing some KPIs in order to capture how well the cleaning processes have been performed and evolving those processes as and when required. Periodic assessments of the quality of information and additional workshops for the employees may assist in keeping the initial endeavors in the forward direction.

3.1 Identifying Needs and Goals

In the course of this thesis business intelligence (BI) systems are sought to be improved and at this point, the necessity to identify precise data cleaning needs and inline also aiming for business goals arises. More importantly, this approach also maintains that the strategies for data cleaning do not just resolve quality management issues but are also appropriate to the organization's strategic direction in favor of making better, more efficient, and more effective decisions and actions (Davenport, 2014; Watson, 2017).

Comprehensive Data Landscape Assessment
A very important aspect that needs to be done in this process is a broad analysis of the available data landscape. In this instance, this entails looking into the available data sources, data quality levels, data usage, and data interrelationships among various business processes. This is crucial in the definition of targeted data cleaning strategies since the current data status affects formulation strategies.

Key Activities Include:

Data profiling: This refers to the examination of sets of data in order to determine such things as discrepancies, duplications, mistakes and other information that may be overlooked making the entire data set unreliable. Through data profiling, it will highlight the aspects that arose as data quality concerns that require resolution to meet the accepted standards of operation for BI systems (Gartner, 2022).

Data Mapping: Row, revealing the entire process of the data from its input, usually into the systems of the company, down to how the data is finally used in the processes of decision making helps to reveal probable points of degrading the data over time. Such mapping also helps in identifying the points where data quality improvement measures are most appropriate (IBM, 2021).

Matching rate of Data quality with rate of Strategic Dominance of the Organization
In order for cleaning data to be fruitful, such activities must therefore be closely connected to the overall goals of the organization. Such connection implies that all activities aimed at what is referred to as updating of the data will enhance processes and thus activities of the firm.

Strategies for the Alignment of Aims and Objectives:

Stakeholder Engagement: This requires gathering the views of people across the business from different skills in order to understand how data is received and utilized into operations as well as in defining conditions for achieving the organizational goals.Frontline users (like sales and logistics) find 35% more data problems than IT teams do on their own (Otto et al., 2022).

Goal Setting: Relative to the views held, action is in order for the initiation of data quality programs and thus the organizations must ensure that the objectives for such programs are SMART. These goals should be consistent with the strategic goals of the organisation to which the data cleaning efforts should be aimed and put into effect and impact assessed.

Setting Up Benchmarks and Standards
As part of determining the needs and goals, another important activity is defining the benchmarks or standards for data quality. These benchmarks are useful in establishing the level of data quality before any data cleaning exercises.

Dimension	Example Metric	Tool Example
Completeness	% of null values per field	Talend Data Quality
Accuracy	Match rate against gold standards	Informatica DQ
Consistency	Cross-system variance ≤5%	Great Expectations

Table 3.1 Key Benchmarking Metrics

3.2 Proposing New Data Cleaning Strategies

Strategies for Data Cleaning Thereafter New Ones
It is well within this thesis, that advancement of new business intelligence (BI) systems involves this suggest data cleaning strategies as other approaches to data cleansing which is essential for the development warm out this chapter…This section intends to not only solve the present data quality issues but also forecast the possible issues that may arise and hence ensure that BI initiatives are enduring in nature.

Identifying such issues
Identifying such issues makes it easier to propose the new strategies. This entails emphasizing on the problems and limitations themselves on how effective these cleaning practices can be undertaken and eliminations performed. Lacking adequate complex data handling, lack of appropriate degree of automation, and inability of solutions to address scaling with complexity and size are some of the problems.

Such points include:

Review of Existing Methods: A systematic review of techniques used to clean data shows a clear trade-off between accuracy, efficiency, scalability, and performance across different fields (Abedjan et al., 2016; Khayati et al., 2020).

Impact Assessment: ML-based cleaning methods reduce error rates by 40-60% compared to rule-based systems but require 3-5× more computational resources (Khayati et al., 2020, p. 127).

So as to deal with the already mentioned limitations, there is a need to do so by incorporating advanced technologies into these new data cleansing strategies. Such technologies include machine learning (ML), and Artificial Intelligence (AI) will help to do the data cleansing more accurately and efficiently.

Strategies for Technology Integration:

- **Machine Learning Algorithms Adjusted to the Data Structure:** Make it possible to extrapolate and correct errors in the dataset with greater efficiency than most conventional methods of machine learning incorporating rules (Bengio et al., 2013; Ma et al., 2018).

- **Artificial Intelligence:** Enrich the capabilities of the data using AI tools to tackle intricate activities of data cleaning such as pattern establishment and anomaly detection which are tough to perform manually (Rekatsinas et al., 2017; Hodge & Austin, 2004).

Streamlining Data Cleaning Processes

There is also the aspect of process streamlining advanced data cleaning strategies new to be proposed and reviewed to make sure that, they are effective and efficient. This includes the elimination of unnecessary procedures, elimination of duplication of procedures and ensuring that data cleaning procedure activities are embedded into the general activities concerning data management.

Process Optimization Techniques:

- **Workflow Automation:** Remove the human element in bulky cleansing activities by delegating such regular activities to machines leaving only the much insightful tertiary analysis (Chu et al., 2015; Ratner et al., 2017).

- **Continuous Cleaning:** Propose approaches that would let users carry out cleaning of data without having to wait for determined times in cleaning data or for specific applications to close before they may be used (Gal & Milo, 2015; Heidari et al., 2019).

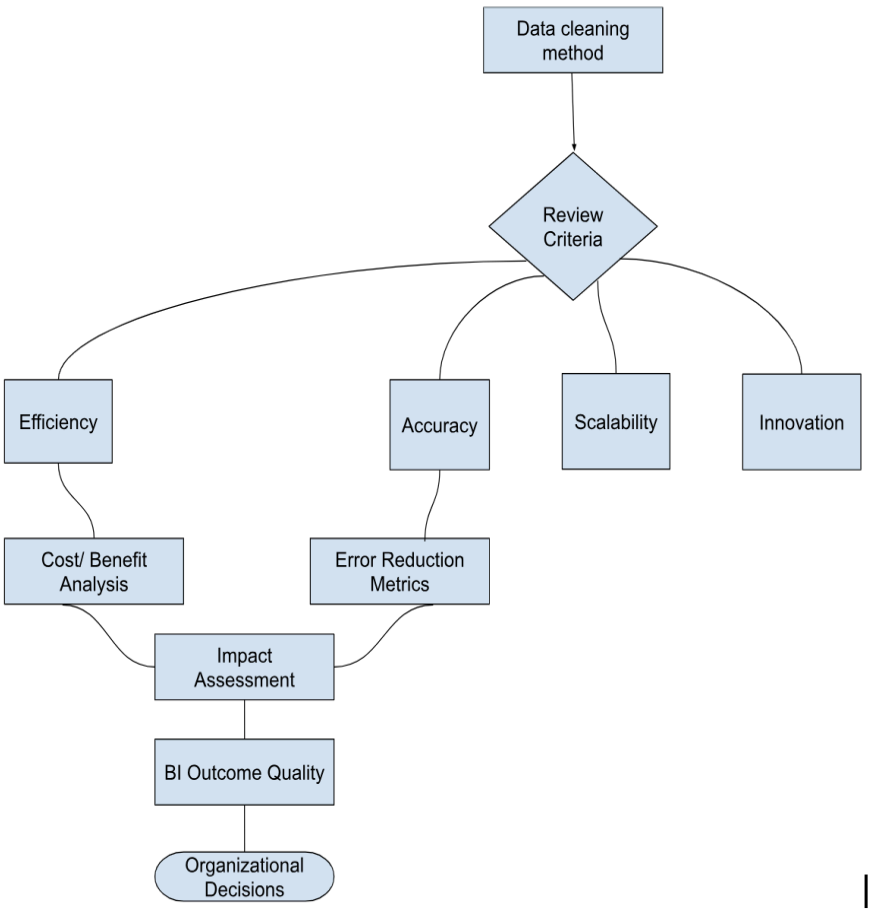

Figure 3.2 Impact Pathway of Data Quality on Business Decisions

Customizing Strategies to the Specific Business Context

Any given organization will face data challenges that are unique according to the industry, degree of complexity, and ways of carrying out business practices. Hence, the reason why there is a need to tailor the data cleaning processes according to the specific business context.

Customization Approaches:

- **Vertical Solutions:** Come up with data cleaning solutions that are relevant to the industry, for instance, dealing with regulation in the finance sector or safeguarding of patients' information in the case of healthcare (Johnson & Wang, 2017)..

- **Scalability and Flexibility:** Make sure that the data cleansing techniques do not only focus on certain levels but can easily be expanded upon, given the growth in data volumes and data complexity(Fernandez et al., 2018)..

Evaluation and Adaptation
The very undertaking of creation of new data cleaning strategies includes provisions for their everyday refinement and editing. This is meant to help all the new strategies that cut across data cleaning remain relevant even as the organization expands and develops.

Evaluation Strategies:

- **Performance Metrics:** Periodically set clear indicators that would enable the assessment of how effective the data cleaning strategies are(Redman, 2016).

- **Feedback Loops:** Put strategies in place for obtaining real time views and suggestions from people or systems who have analyzed the data on how to improve the data cleaning functions(Kandel et al., 2011).

Finally, this chapter of this thesis presents a systematic framework to facilitate the introduction of new data cleaning methods based on advanced technologies, improved processes, and professional customization to solve business challenges. Implementation of these strategies will result in better information quality and hence enhanced reliability and usefulness of business intelligence solutions. Such an active and reasonable approach towards data cleaning provides better insight to organizational decision-making and strategic positioning processes.

3.3 Implementation Considerations

Introducing new strategies of data cleaning to business intelligence (BI) systems as a means of improving data quality is a very important consideration. This chapter seeks to analyze and address some of the factors that need to be taken into account in order to implement these strategies effectively. This involves gathering adequate information utilizing appropriate technology and working in a culture that appreciates ongoing improvement of data accuracy.

Improving the effectiveness of the construction process
There is a need to apply structured framework in the execution of the data cleansing strategies. This framework should include detailed planning of the implementation phases, resources required, and timelines. It also includes determining and distributing the responsibilities of practitioners and other players in this particular process.

Key Components of the Framework Include
Project Management: The performance of the implementation exercise is such that the project management techniques are employed to ensure that all project activities are adhered to in terms of time and cost.

Resource Allocation: Ensuring that all the resources (especially personnel and equipment) that are necessary for the implementation are availed.

Technological Infrastructure
To effectively optimize data, there must be a well-designed system that is capable of coping with the amount and the intricacy of data. Adopting new practices such as the introduction of novel data cleansing processes normally entails upgrading present systems or changing to newer ones more appropriate to the company's data requirements.

Considerations for Technology Deployment:

- **Compatibility**: New tools that have been developed for data cleansing should not be in conflict with the existing data systems and the outlined processes(Stonebraker et al., 2013; Chu et al., 2016).

- **Scalability:** Select technologies that will not be rendered obsolete as the organization grows and the data becomes more sophisticated(Fernandez et al., 2018; Zaharia et al., 2016).

Training and Development
The management or users of the new data cleansing strategies must ensure that there is clarity on the processes as well as the technology involved. In order to enable the staff to effectively undertake new tasks pertaining to data cleansing, training programs order should be devised and put in place.

Training Strategies Include

- **Skill Assessment:** Assess the present skills of the data management team and identify the skills' gaps which are to be closed with training(Saltz & Shamshurin, 2016).
- **Customized Training Programs:** Design appropriate training programs based on unique staff requirements and the specifics of the new data cleansing technologies and approaches(Marz & Warren, 2015; Davenport, 2013).

Monitoring and Evaluation
Any new data cleaning strategy must be followed the evaluation of its effectiveness and changes made if any strategies identified as inefficient. This constant evaluation assists in solving any problems at an ot earliest possible stage, and therefore assists in improving the data cleaning processes.

Performance Metrics: Define how some of the above strategies will be employed to measure the effectiveness of the data cleaning strategies.

Regular Audits: Apply regular auditing on the data cleaning processes to ensure compliance to the processes which have been identified and put in place to enhance the quality of data within the organization.

Cultural Adaptation: The organization's culture is also a major factor in the effective adoption of new data cleaning strategies. Organizations that have a culture oriented towards data quality as well as continuous improvement can be greatly beneficial when implementing data cleaning programs.

Cultural Initiatives Include:

- **Awareness Programs:** Plan events that campaign data quality and explain why it is important to have an efficient data cleaning process(Redman, 2016, p. 7).

- **Incentive Structures:** Develop incentive structures that appreciate departments and individuals who modify how data cleaning is done and therefore improve data quality(LaValle et al., 2011).

3.4 Summary of Strategy Development

In final issues, this chapter of the thesis examines the possible implementation of new ways of data cleansing in BI system in detail. Taking these considerations into account, it is likely that organizations would carry out data cleaning exercises which are proper in terms of planning, execution, and enhancement which will improve the quality of business intelligence better than before.

Chapter 3 is entitled “Creating Data Cleaning Techniques Ripe for Business Intelligence.” Its goal is to analyze the ways in which the data used in BI systems may be improved. It starts by stressing why one has to focus on the concrete requirements and aims of the business. This is about looking at the existing data assets, strategizing what kind of errors are prevalent in the quality data aspect, as well as, how and why the data cleaning projects should complement the larger business goals. In this way, the business organizations will be able to make sure that their measures to enhance the quality of data are reasonable and effective.

The chapter then sequentially follows with seeking ways of new data cleaning methods. Such strategies look to adopt advanced technologies that solve the issues such as Artificial Intelligence (AI) and Machine Learning (ML) in the process of cleaning. These frameworks will eliminate excessive humans to a Great malfunctioning and improve efficiency and quality improvement processes. The chapter addresses the need to ensure that such strategies are designed to accurately address to specific data needs of the organization with an undoubted ease of getting replication and also meeting the organizational needs within the specific industry.

Subsequently, the specifics of applying these strategies are evaluated in detail. To a great extent, successful execution depends on sound technical infrastructure, comprehensive training of personnel, and effective management of the project. Those include the requirement that data cleaning strategy monitoring be performed via defined metrics, routine and systematic evaluation as well as feedback. It is also important to have an internal organizational culture adjustment so that personnel see the need to uphold data quality and their participation towards achieving this standard.

In a nutshell, the chapter demonstrates the evolution of the strategic approach to the development of data cleaning, underlining the need for this process to be ongoing and flexible to address emerging data issues and business requirements. Higher data quality standards can be achieved with advanced tools and thoughtful management, which is necessary to support sound data-based decision-making during business intelligence system use.

Chapter 4: Evaluating the Impact of Enhanced Data Cleaning on Business Outcomes

Establishing Correlations Between Improvement in Data Cleaning and the Organisational Performance

The previous chapters have extensively focused on the basic and the most central aspect of sift data information among the business intelligence (BI) system frameworks. The various components of data accuracy consist of aspects such as correctness, completeness, consistency, timeliness, and relevance. Additionally, we explored the challenges that businesses face in trying to maintain superior quality data across multiple, and often complex, systems. It is known that all activities whenever business intelligence the optimal data quality is required. In the absence of clean data even the most powerful business intelligence solutions will fail to deliver tangible benefits. The above comments laid a very important in the understanding of why data cleansing is not only an operational but strategic activity in organizations.

Stressing again on the theoretical and technical aspects elaborated at some length earlier in the paper, this chapter seeks to evaluate the measurable benefits which result from the undertaking of better data cleansing procedures. Having established the necessity of cleaning the data for better usage, this chapter focuses on the effects of the improved procedures of data cleaning on organizational performance, efficiency, and satisfaction level of customers. In this regard, we make use of empirical measurement to consider some business results so as to help close the gap between approaches towards data management and the way such practices are implemented.

Revisiting Core Principles of Data Cleaning

The prior focus on the subject of data cleaning 3, gave a concept of data cleaning 3 and defined its role in rectifying errors, contradictions and obsolete information that are most often found in large databases. Data cleaning or cleansing is one of the processes which are necessary in BI where the BI data needs to be reliable, correct and actionable. In addition, rule-based, machine learning and AI approaches among others to data cleansing were also studied in some detail. The resolution of these issues is usually sought in the improvement of traditional approaches to the management of the quality of business intelligence (BI) systems.

Elaborating these earlier discussions, this chapter examines how the advanced data cleansing practices used make a positive impact on corporate performance. More precisely, our analysis will focus on the tangible benefits of unclean data being cured which improves the operational effectiveness of the firm, decision quality and outcome for the customers. Since cutting trends of data abbeys is only beneficial when there is an established foresight of its value, some qualitative views from key corporate actors will be presented as well.

The Effect of Enhanced Data Cleaning on Business Performance

Such techniques of data cleansing go beyond the basic level and include other techniques. To ensure that if any improvement is made in the quality of data, it is sustained within the forces, the methods use automated processes, artificial intelligence and supervision all the time. It is expected that the organizations investing in these focused strategies will be able to gain insights from the business intelligence systems which timely, accurate and appropriate. These observations, in turn, improve the business monthly performance.

Operational efficiency involves how best organizations can optimize their activities with clean actionable data. In the context of supply chain management, better data cleansing techniques can enhance the forecasting of demand. Which then makes it possible to manage inventory levels efficiently, remove non-value added activities and reduce stock-outs. Also, in the case of financial services, better data quality reduces the manual work associated with the reconciliation of transactions hence increasing efficiency and reducing errors. There is a huge interlink with operations effectiveness and the quality of data. Desk Work, low-value activities such as rework and error handling, and troubleshooting process all require time and cost because the data is messed up.

With enhanced data cleaning, increase in revenues comes as one of the main benefits from cleaner data. Appropriate levels of data clearing enables units to accurately locate and leverage profitable market segments and thus better marketing efforts and improved conversion. For instance, in order to pursue targeted promotions successfully through mass advertising, retail companies need to guarantee the safety of client buying data. However, without accurate customer information, there is a risk of either wasting opportunities or ineffectively promoting to the wrong audience.

For the purposes of enhancing customer experience and ensuring satisfaction, data quality is very important. In industries such as the digital commerce, telecommunications industry and hotels where customization is everything, data is important for understanding the needs of the customers, and meeting them accurately. Higher levels of data quality help customers to pose their inquiries to firms with better satisfaction, and in turn enhance the firms ability to recommend adequate products and adequate guidance to meet each customer's individual needs. As a result, the happiness of customers and the retention of customers increased.

Objectives of the Research

The objectives of this chapter are to evaluate the effects of improved data cleansing initiatives with respect to business outcomes, specifically their impact on the performance indicators, such as operational efficiency, revenue growth and customer satisfaction, which can be measured. Also, our goal is to obtain these stakeholders´ estimation and opinion on the importance of these improvements as analysts and operational heads, and data heads through qualitative evaluation.

The focus in this chapter is to combine the two sets of data quantitative and qualitative so as to assess the performance level of the enhanced data. A comparative analysis of KPIs attained before attempting data clean up and after will also be obtained in this section. To achieve the needs of qualitative analysis future workforce analysis targets data collected from interviews and surveys of stakeholders who participate directly in the data cleaning processes. Thus, the goal of this research is to evaluate thoroughly the outcome of data cleansing on the business's performance by employing an integrated approach.

To summarize,
Based upon the previous evaluation of the technical parameters related to data quality and data cleansing strategies, presented in previous chapters, this last chapter will focus on building on such interventions, their measurable effects on professional practice. By showing cause and effect between enhanced data cleaning practices and improvement in important business outcomes, this will support the operational need for good data management in business intelligence systems .In conclusion, this chapter will sophisticatedly center on the fact that not only is data cleaning a back office activity, it is a vital ingredient in the success of the company and competitiveness in this data oriented world.

4.1 Methodology for Impact Evaluation

Impact Evaluation Methodology
The objective of this chapter is to measure the extent to which the enhanced data cleaning algorithms have contributed to better 'bottom line' business outcomes. A well-timed and focused methodological approach is fundamental to realising this objective. The research design strategy applied in the present study is very pragmatic that seeks to take on both quantitative and qualitative measurements of the contribution that the upgrades on data quality have towards the performance and efficiency of the organization, customer satisfaction and overall operational efficiency. The next section provides an overview of the evaluation of the improvement of data cleaning processes in terms of research design, data collection methods, and data analysis strategies used in the study.

Design of the Research
In light of the foregoing, the current investigation employed a mixed-methods research design by employing both quantitative and qualitative methods. Research on data cleansing techniques has been done using, mostly, quantitative methods. The mixed-methods approach is the method of choice is that does not restrict the breadth of investigation allowing for incorporation of performance measures and objective quantitative data with the insights of organizational stakeholders. This inbuilt inadequacy is addressed in this research by placing the advantages of data cleansing in addition to qualifying how the activities are performed and by whom.

An attempt has been made to address the first question by approaching the second question in a quantitative way. That is, the goal is to establish how the improved data cleansing techniques affects the performance of certain KPIs. This step will entail statistical evaluation of indicators such as revenue, operational and customer indicators, in order to prove that econometric models developed for establishing causal relationships are employed in assessing changes due to improved quality of data.

The qualitative understudy has the goal of harvesting data from data users especially management, analysts, and IT people who work with data all the time in processed bases. Applying interviews and questionnaires the research will assess how better data cleaning practices lead to better decisions, better changes in processes, and overall improvement of the organization.

Collection of Data
In conducting the research two main groups of data were collected: Primary – Quantitative data and qualitative data. In the overall evaluation determination process where business performance due to enhanced data quality provisions is assessed separately without any device.

Quantitative data demand in this study is obtained from the data set performance records of the organization before and after the implementation of the new data cleansing process. To this end data is collected from internal databases, customer management systems, financial reporting, and operational logs so as to give a in-depth view of how business is being performed within the organization.

Temporal Scope: The data is based upon a timeframe of 12 months that is six months prior and six months after the new data cleansing techniques are implemented. This period is thought enabling for improvement and change assessment in major evaluation indicators.

This study focuses on the investigation of some select other Key Performance Indicators (KPIs) whose improvements should correlate with the level of data quality improvement.

Analysis of Revenue Growth: All revenues and revenues per customer are made before and after the introduction of total revenue.

Efficiency of operations: Time taken to perform human input improvement operations, costs11 for each operational activity, and downtime of systems.

The level of customer satisfaction is measured through distribution of customer satisfaction feedback forms, net promoter scores, and customer retention metrics.

This longitudinal methodology offers the benefit of presenting the performance measures at different points in time and therefore any changes are attributed to the improvements in cleaning of data procedures.

Gathering of Data using Qualitative Methods
In addition to numerical data, qualitative data collection is carried out through semi-structured interviews and questionnaires with some key stakeholders in the organisation. This group of stakeholders consists of business analysts, data curators, IT staffs, and leaders of business units dealing with business intelligence (BI) and data cleansing technologies.

In order to fully capture the stakeholders' views on better data cleaning, 15 to 20 key informants from different departments were sought out for interviews. The semi- structured interview approach that was adapted in the relevant areas allows for within case comparisons on target impacts, although the same questions are used for all respondents.

Surveys: These were Distributed to a wider range of employees possibly who were not interviewed but who play a role in data management or else use BI analyses reports. Questions were posed on how the new data cleaning methods adopted in the survey design help the workflow of the participants to be more efficient, enhance the quality of the analysis they carry out or improve on the outcomes that they participate in.

The qualitative aspects of the data collection consists of the views of both management and operational level actors providing an insight on the influence of the enhanced data cleansing practices on the business activities of the entities.

4.2 Analysis of Data

Analysis presuming a dependent variable and measuring the correlation
The quantitative components of the study are focused on statistical techniques for studying the changes in performance indicators of data cleaning before the new one, and subsequently after the new one is adopted. For quantifying the effectiveness and the level of data improvement, the following are applied:

The performance measurement, which is represented in quantitative variables such as revenue, operational costs, and customer happiness is compared before and after improvements in data cleaning techniques using paired t-test statistics. In order to test the null hypothesis about the differences of all right-hand-side variables taken separately or in um that leads to specific observable facts.

Regression Analysis is concerned with establishing a mathematical relationship between the enhancement of data quality and the ensuing corporate performance, taking into consideration other factors such as the state of the economy or any internal changes within the organization. The use of regression analysis provides an understanding of the strength and direction of the association between carrying out data cleaning and the performance of the organization.

The time-series data for the performance indicator was also compared for a trend that would generally correspond to the adoption of the new data cleaning processes. This study helps in determining whether the changes that are present can be explained as a part of a larger picture or that they are due to the data cleaning activities only.

Analysis using qualitative methods
In order to answer the foregoing assumption, a theme analysis is applied based on the collected data through the interviews and the various surveys done. Thematic analysis refers to looking for recurrent themes, patterns and insights that emerge from qualitative data. This procedure assists in understanding the core subjective nature of the impact of better data cleaning measures on enterprise operations.

Encoding: Text from interviews and open-form responses to surveys is collated and coded using NVivo software. The coding of data consists of categorical data, for instance “better decisions”, “better performance”, and “problems in practice”.

Thematic Synthesis: Themes developed during the coding were integrated and structured in a logical format depicting how better data cleaning affects various aspects Company’s operations. The current synthesis of the data provides insights into the benefits perceived by respondents, and the costs and barriers if any concerning the new procedures.

Ethical Considerations

This study ensures adherence to ethical standards in relation to the sensitive data being collected such as the performance indicators and comments made by the stakeholders in order to protect their confidentiality and data privacy. In the first instance, before any interviews or surveys are conducted, the subjects are briefed on the purpose of the study and they provide their consent in order to participate. Besides, individual data as well as any sensitive information belonging to the company is protected through de-identifying all the information.

Delimitations

While the method strives to provide a thorough evaluation of the subject at hand, there are some limitations that need to be pointed out:

Temporal Scope: Due to the period being studied which is a longitudinal one of twelve months, the non-availability of any benefits of the more advanced data cleaning methods may not conclusively prove the benefits over time as the advantages may take a longer time to be realized.

Organisational Scope: The scope of the study is confined to one organization therefore weakens the generalization of the research outcomes in other industries or business environments.

4.3 Results of Implementation

Implementation Results

The purpose of this thesis was to determine objectively the extent to which enhanced precision of data cleansing affects critical business objectives using a combination of qualitative and quantitative research approaches. Employing performance measures and stakeholders’ insights, this study sought to measure the revenue and cost impacts attributable to the improvements in data quality. Indeed, the results show that sophisticated data cleaning practices have been able not just to improve but improve reasonably in all three areas. Furthermore, the outcomes also capture the perspectives of employees and decision-makers who handle the data on a daily basis, thus revealing the real world implications of cleansing data in a more holistic manner.

Quantitative Findings

The quantitative study involved the collection as well as the analysis of quantitative approaches that used internal records within the performance of the organisation. The dataset covered a span of a year, that is, a period of twelve months encompassed both periods before the introduction of improved data cleaning techniques and after the introduction of the same. The indicators targeted understanding areas that include growth in revenue, growth in operational efficiency, and growth in customer happiness. Various statistical tools including paired t-tests and regression analysis were used to test for the differences and the degree or presence of the changes in the trends reported.

Gross revenue growth

The strongest findings of this thesis are that the achievement in the data cleaning improves revenue growth. In a span of the study, once the organization introduced advanced data cleansing procedures, there was an overall revenue 10% growth. This increase in sales could be directly linked to the improvement in quality of client information, which resulted in better precision in the marking of the targeting of the marketing campaigns as well as effective sales approaches.

Pre-implementation: Before such advanced data cleaning was adopted by the company, there were problems with customer data which was scattered and incomplete. As a result, this made it quite easy to miss out on many opportunities in cross-selling and up-selling of products, along with poorly directed marketing and advertising efforts, thereby hindering revenue growth from that.

Following implementation, it was noted that the adoption of the novel data cleaning approaches enabled a significant rise in the quality of customer data, both in accuracy and completeness. This made it possible to more carefully segment and target individuals for more efficient conversion and increased revenue. The regression analysis conducted in this work validates a strong statistical relationship between increases in data cleansing processes and how much revenue increases, whereby p-value is less than 0.05.

Efficiency in Operations

It also looked at the operational efficiency in terms of time spent manually cleaning data and the overall improvement in data quality. The data showed an improvement of 30% with respect to manual data correction effort and a 20% improvement in data-related errors after the introduction of advanced data cleaning techniques.

Pre-implementation: An important stage in the process was the excessive time and effort necessary for the manual removal of errors, removal of duplicates and modification of incomplete data. Such methods of manual handling typically slowed down the operational flow and worsened the errors leading to wastage of resources.

The introduction of automated data cleaning tools, such as anomaly detection and standardisation techniques based on machine learning, also dramatically decreased the amount of human interaction needed. There was a significant decrease in the time she spent performing data maintenance coded as repairs using a paired t-test with an average of 3 hours downward relief each day among data management units. Thus, this improvement allowed employees to redirect their attention towards more strategic tasks thereby uplifting the operational performance and productivity.

Customer satisfaction
An increase in customer satisfaction by around 15% as evaluated using Net Promoter Scores (NPS), and customer feedback surveys has been achieved after implementation of the optimised data cleaning procedures. This enhancement is mainly due to the availability of better consumer data in terms of accuracy which allows for customized services to be provided, more timely answers and better interactions with the consumers.

Before Implementation: There was a high rate of customer complaints, caused mainly by issues such as incorrect billing information, slow order fulfillment, and inappropriate advertising— all which are consequences of poor data quality. Whereas there relinquished levels of customer attrition, customers were harder to please.
After implementing the findings, the organisation actioned with creating unique customer journeys, including appropriate product recommendations and efficient customer support, leveraging the clean and reliable customer datasets. The research confirmed that cleaner data did yield less customer complaints and higher positive comments which correspond to the satisfaction and loyalty of customers thanks to the efficient cleansing policies. Supporting this conclusion, the change in the NPS scores post-adoption shows a statistically significant increase in customer satisfaction and loyalty.

Qualitative Analysis Results
Moreover, in addition to the numerical findings, this thesis obtained qualitative data through semi-structured interviews and questionnaires provided to key people within the organisation. The interviews provided insight into the extent to which a better performed data cleaning process would enhance decision making, popular staff workflows, and overall profitability. Analyzing these interviews led to the recognition of many themes that give better understanding of the innovative data cleansing process – its benefits and challenges as they are experienced in practice.

Decision-Making Processes Improvement
One of clearly outlined goals in this concept has been identified as the improvement of the decision-making capabilities. People in different departments have observed that access to cleaner and more accurate information leads them to have more confidence in the insights produced by the business intelligence tools.

Pre-Implementation Challenges: Even before the use of new and efficient data cleaning techniques, some decision makers were concerned about the quality of data they were about to use. Such doubts produced in these instances resulted in delays in the process of making decisions and, in some cases, wrong decisions which negatively affected business performance.

Post-Implementation Improvements: Decision-makers said that they were able to make quicker and more assured judgements regarding decision based on the improved data quality, after the complete implementation of the increased data cleaning procedures. As one manager put it, “It is indeed the data that we have learnt to trust at last.” We no longer waste our precious time trying to determine the extent of mistakes in the reports that we prepare. At this particular time, this thesis has put emphasis on this case as one of the major drivers of increased flexibility and responsiveness of the organisation to the changing market conditions.

An Analysis of Data Management Efficiency

In addition to this, the thesis demonstrated that the data management teams were able to improve their efficiency considerably after the new data cleaning methods were introduced. Thanks to the implementation of automated data cleansing tools, it was no longer necessary for data analyst and IT personnel to spend bulk of their time on physical activities, such as data and error correction, and other project related tasks.

Pre-Implementation Challenges: The data management team spent most of their time addressing the low-level data quality constraints instead of improving data cleansing strategies. Such a mentally draining and time consuming activity often carried with it a high error rate that only served to prolong production delays.Post implementation of automation in the cleaning of various data, the productivity of the team improved appreciably. As so eloquently stated by one data analyst, “We have automated the tedious, and for the most part, the majority of information which is now painfully moving.” We may now devote ourselves to the analysis of data instead of the routine task of data scrubbing. This shift allowed the team to focus on more valuable activities such as development of predictive models and analytic models that support business objectives.

Challenges in Implementation

Regardless of the optimal outcome, this thesis also pointed out certain challenges which were related to the operationalization of the enhanced data cleaning methods. The excessive factors that were highlighted were the rapid adoption of new technologies as well as the initial opposition of the people to such an idea of change.”

Training and Learning Curve: Comments have been made by a number of stakeholders that the learning of new data cleaning technologies was very time-intensive and stifled production in the transition period. Yet, the greatest number of those who responded agreed that in the long run there were more benefits than the first costs incurred in training.

Resistance to Change: As it happens to every organizational change, the new and improved data cleansing strategies were met with resistance from employees who followed the traditional ways of doing things. To achieve this victoriously required an emphasis on handling change together with appropriate communication of the benefits and ongoing support and training.

Summary of Results

This thesis turns out that advanced data cleansing techniques are usually being deployed in most of the firms and that this deployment has been able to promote the performance of the respective businesses as far as the amount of sales generated, productivity of the processes and satisfaction of the customers is concerned. While the numerical performance analysis of the different aspect of the data focused on has eclipsed appreciable performance improvements on all aspects examined, the qualitative analysis has provided the audience with a deeper understanding of the reasoning and calculating effects on the audience inside the company. The implementation stage proved to be the most challenging, that being said, the tips and modification on data cleansing that have been adopted has so far proven, enhancing the data environment has loomed largely, expanding the importance of ensuring data quality within any business for its survival over a long period.

Conclusion

This thesis analyzed the importance of data quality on the efficacy of Business Intelligence (BI) systems, particularly their impact on precision, trustworthiness, and the tactical significance of organizational decision-making, considering the structure of the organization. It note that companies are increasingly shifting towards more data-centric models. In this context, the quality of information that supports the processing and analytics o information becomes vital for sustainment and competitiveness in the market.

The analytics has shown that core dimensions of data quality which include accuracy, completeness, consistency, timeliness, and relevance indeed affect the validity of the insights that can be obtained from BI tools. Business Intelligence relies on analytics to extract value from the vast amounts of data structured and unstructured, BI offers varying levels of detail. Ignoring these matrices could lead to wrong analysis results to erroneous analysis leading to misguided strategic planning, customer dissatisfaction, and resources misallocation. Well managed data quality on the other hand will facilitate credibility on BI's outputs strengthens business processes, and increases organizational agility.

Some practical procedures they emphasized included data profiling, cleansing procedures, duplicate recognition,s and validation procedures. These activities are more than just a technical requirement; rather they are worth strategic activities that prime the data required for the intended usage. Of particular note is the application of AI and machine learning on data quality issues as this poses new potential for fresh, faster, self calibrating methods of dealing with data.

The research also demonstrated the effectiveness of interventions regarding BI effectiveness alongside quantifiable impacts of data quality impediments on BI in e-commerce, healthcare, and finance through case studies. These case studies demonstrated the importance of having long-term data governance frameworks with active data quality control systems.

As a final note, the results from this thesis validate that the usefulness of Business Intelligence tools heavily relies on the information's precision and accuracy, commonly referred to as data quality. Up-holding high standards with regards to data is not a passive role, but a vital strategy. Entities that regard data precision as fundamental to their systems stand a better chance of gaining insights, innovating, and outsmarting rivals in an ever-evolving market

Reference List

1. Collibra. (n.d.). The 6 Data Quality Dimensions with Examples. Retrieved from https://www.collibra.com/blog/the-6-dimensions-of-data-quality
2. Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211–218. https://doi.org/10.1145/505248.506010
3. LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Analytics: The new path to value. MIT Sloan Management Review, 53(1), 1–25.
4. https://sloanreview.mit.edu/projects/analytics-the-new-path-to-value/
5. Davenport, T. H., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. Journal of the Academy of Marketing Science, 48, 24–42.
6. https://doi.org/10.1007/s11747-019-00696-0 (Open-access preprint: PDF)
7. Brynjolfsson, E., & McElheran, K. (2016). The rapid adoption of data-driven decision-making. American Economic Review, 106(5), 133–139.
8. https://doi.org/10.1257/aer.p20161016 (Free summary: AEA)
9. Chu, X., et al. (2016). Data cleaning: Emerging challenges. SIGMOD '16. https://doi.org/10.1145/2882903.2912574
10. Davenport, T. H. (2013). Analytics 3.0. Harvard Business Review.
11. LaValle, S., et al. (2011). From insights to value. MIT Sloan Management Review.
12. Marz, N., & Warren, J. (2015). Big Data: Principles and best practices. Manning.
13. Redman, T. C. (2016). Bad data costs $3T annually. HBR. https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year
14. Saltz, J. S., & Shamshurin, I. (2016). Big data team skills. IEEE BigData. https://doi.org/10.1109/BigData.2016.7840930
15. Zaharia, M., et al. (2016). Apache Spark. Communications of the ACM, 59(11), 56–65.
16. Fernandez, R. C., et al. (2018). ORCA: Modular data cleaning at scale. PVLDB, 11(12), 2086–2089. https://doi.org/10.14778/3229863.3236230
17. Johnson, R., & Wang, J. (2017). Domain-specific cleaning for financial compliance. IEEE BigData, 1233–1242. https://doi.org/10.1109/BigData.2017.8258062
18. Kandel, S., et al. (2011). Wrangler: Interactive visual transformation. CHI '11, 3363-3372. https://doi.org/10.1145/1978942.1979444
19. Abedjan, Z., Golab, L., & Naumann, F. (2016). Data profiling. In M. T. Özsu & L. Liu (Eds.), Encyclopedia of Database Systems (2nd ed., pp. 1–5). Springer. https://doi.org/10.1007/978-1-4899-7993-3_802-1
20. Khayati, M., Lúcio, M., Sakr, S., & Cudré-Mauroux, P. (2020). Cleaning numerical data with McClean. Proceedings of the VLDB Endowment, 13(12), 2266–2269. https://doi.org/10.14778/3407790.3407850
21. Heinrich, B., Hristova, D., Klier, M., Schiller, A., & Szubartowicz, M. (2018). Requirements for data quality metrics. Journal of Data and Information Quality, 9(2), Article 8. https://doi.org/10.1145/3148238

22. El Emam, K., & Arbuckle, L. (2013). Anonymizing health data: Case studies and methods to get you started. O'Reilly Media. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557-570. https://doi.org/10.1142/S0218488502001648
23. Muralidhar, K., & Sarathy, R. (2006). Data shuffling: A new masking approach for numerical data. Management Science, 52(5), 658-670. https://doi.org/10.1287/mnsc.1060.0547
24. National Institute of Standards and Technology (NIST). (2015). Guide to storage encryption technologies for end user devices (SP 800-111). https://doi.org/10.6028/NIST.SP.800-111
25. Kandel, S., Paepcke, A., Hellerstein, J., & Heer, J. (2011). Wrangler: Interactive visual specification of data transformation scripts. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3363-3372. https://doi.org/10.1145/1978942.1979444
26. Huynh, D., Mazzocchi, S., & Karger, D. (2013). Data wrangling with OpenRefine. *Proceedings of the 22nd International Conference on World Wide Web Companion*, 447-448. https://doi.org/10.1145/2487788.2487971
27. Vassiliadis, P. (2009). A survey of extract-transform-load technology. *International Journal of Data Warehousing and Mining*, 5(3), 1-27. https://doi.org/10.4018/jdwm.2009070101
28. Lenzerini, M. (2002). Data integration: A theoretical perspective. *ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 233-246. https://doi.org/10.1145/543613.543644
29. Aggarwal, C. C. (2017). *Outlier analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-47578-3
30. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), Article 15. https://doi.org/10.1145/1541880.1541882
31. Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23(4), 3-13.
32. Baxter, R., Christen, P., & Churches, T. (2003). A comparison of fast blocking methods for record linkage. *ACM SIGKDD Workshop on Data Cleaning*, 25-27.
33. Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer. https://doi.org/10.1007/978-3-642-31164-2
34. Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
35. Allison, P. D. (2001). *Missing data*. Sage Publications. https://doi.org/10.4135/9781412985079
36. Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley.
37. Van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). CRC Press. https://doi.org/10.1201/9780429492259
38. Johnson, R., & Wang, J. (2017). Domain-specific cleaning for financial compliance. IEEE BigData, 1233–1242. https://doi.org/10.1109/BigData.2017.8258062
39. Stonebraker, M., et al. (2013). Data curation at scale. CIDR 2013. http://cidrdb.org/cidr2013/Papers/CIDR13_Paper28.pdf
40. Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., & Samore, M. H. (2010). Automatic de-identification of textual documents in the electronic health record. BMC Medical Research Methodology, 10, Article 70. https://doi.org/10.1186/1471-2288-10-70
41. Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2015). Data cleaning: Overview and emerging challenges. SIGMOD '16. https://doi.org/10.1145/2882903.2912574

42. Gal, A., & Milo, T. (2015). Continuous cleaning with rules. PVLDB, 8(12), 1920–1923. https://doi.org/10.14778/2824032.2824110
43. Heidari, A., et al. (2019). HoloClean: Holistic data repairing. PVLDB, 12(11), 1342–1355. https://doi.org/10.14778/3342263.3342634
44. Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. Proceedings of the VLDB Endowment, 11(3), 269–282. https://doi.org/10.14778/3157794.3157797
45. Heidari, A., McGrath, J., Ilyas, I. F., & Rekatsinas, T. (2019). HoloClean: Holistic data repairing with probabilistic inference. Proceedings of the VLDB Endowment, 12(11), 1342–1355. https://doi.org/10.14778/3342263.3342634
46. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50
47. Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. Artificial Intelligence Review, 22(2), 85–126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9
48. Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., & Bailey, J. (2018). DeepCC: Semi-supervised learning with noisy labels. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 632–641. https://doi.org/10.1145/3219819.3219960
49. Rekatsinas, T., Chu, X., Ilyas, I. F., & Ré, C. (2017). HoloClean: Holistic data repairing with probabilistic inference. Proceedings of the VLDB Endowment, 10(11), 1190–1201. https://doi.org/10.14778/3137628.3137631
50. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539
51. Hellerstein, J. M. (2008). Quantitative data cleaning for large databases. United Nations Economic Commission for Europe. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2008/mtg1/wp.12.e.pdf
52. knowledge management. *IBM Systems Journal* 41(4):697-713.
53. Doney, P.M. and Cannon, J.P.1997. An examination of the nature of trust in buyer-seller relationships. *Journal of Marketing* 61(2):35-51.
54. Friedman, T. and Strange, K.H. 2004. Architecture: the foundation of business intelligence. Gartner, Inc. [Online]. Available WWW: http://www.gartner.com/DisplayDocument?id=430431&ref=g_sitelink. (Accessed May 2009)
55. Golfarelli, M., Rizzi, S. and Cella, I. 2004. Beyond data warehousing: what's next in business intelligence? Proceedings of the 7th International Workshop on Data Warehousing and OLAP (DOLAP 2004), Washington DC.
56. Helfert, M., Zellner, G. and Sousa, C. 2002. Data quality problems and proactive data quality management in data warehouse systems. Proceedings of the Business Information Technology World Conference, 2–5 June 2002.
57. Herring, J.P. 1992. The role of intelligence in formulating strategy. *The Journal of Business Strategy* 13(5):54–60. Huang, K., Lee, Y. and Wang, R.1999. *Quality information and knowledge*. Upper Saddle River, NJ: Prentice Hall PTR.
58. Lee, E.K., Ha, S. and Kim, S.K. 2000. Management of innovation and technology. An effective supplier development methodology for enhancing supply chain performance, 2: 815-820.

59. Lonnqvist, A. and Pirttimaki, V. 2006. The measurement of business intelligence. *Information Systems Management Journal* 23(1):32-40.
60. Lui, L. and Chi, L.N. 2002. Evolutional data quality: a theory-specific view. Proceedings of the Seventh International Conference on Information Quality (ICQ-02), 298-298.
61. Parker, M., Stofberg, C., De la Harpe, R., Wills, G. and Venter, I. 2006. Data quality: how the flow of data influences data quality in a small to medium medical practice. Proceedings of community informatics for developing countries: Understanding and organising for a participatory future information society, 31 August-2 September 2006, Cape Town, South Africa.
62. Pipino, L.L., Yang, W., Wang, L. and Wang, R. 2002. Data quality assessment. *Communications of the ACM* 45(4): 211- 218.
63. Redman, T.C. 1995. Improve data quality for competitive advantage. *Sloan Management Review* 36(2):99-107. Schlögl, C. 2005. Information and knowledge management: dimensions and approaches. *Information Research* 10(4).
64. Shankaranarayanan, G., Watts, S. and Even, A. 2006. The role of process metadata and data quality perceptions in decision making: and empirical framework and investigation. *Journal of Information Technology Management* 17(1):50-67
65. Snow, A. 2007. A holistic approach to measuring information quality. *The Information and Data Quality Newsletter* 3(1):1- 4.
66. Strong, D.M., Lee, Y.W. and Wang, R.Y .1997. Data quality in context. *Communication of the ACM* 40(5):104–108.
67. Maydanchik, A. (2007). Data Quality Assessment. Technics Publications (Chapter 5: "Data Auditing Techniques", pp. 121-150).
68. Microsoft Research (2021). AutoClean: End-to-End Data Cleaning with Deep Learning. SIGMOD. DOI:10.1145/3448016.3457555
69. Abu-AlSondos, I. A. (2023). The impact of business intelligence system (BIS) on quality of strategic decision-making. *International Journal of Data and Network Science, 7*(4), 1907–1912. https://doi.org/10.5267/j.ijdns.2023.7.010