Data Quality

How to Manage your Business Data

Rupa Mahanti, PhD,

Business and Information Management Consultant and Author,

Associate Editor, *Software Quality Professional*

INTELEX

Objectives

You will learn about:

1. Data and the **evolution** of data
2. Why is Data Quality **important**
3. Data Quality **Dimensions**
4. Data Quality **Management**

INTELEX

1: Data and the Evolution of Data

What are Data?

Oxford Dictionary Definition - Facts and statistics collected together for reference or analysis.

IT Perspective - Data are abstract representations of selected features of real-world entities, events and concepts, expressed and understood through clearly definable conventions related to their meaning, format, collection and storage.

Sebastian-Coleman, Laura, December 31, 2012 Measuring Data Quality for Ongoing Improvement, Morgan Kaufmann Print ISBN-13: 978-0- 12-397033- 6
Mahanti, Rupa. (2019). Data Quality: Dimensions, Measurement, Strategy, Management and Governance. ASQ Quality Press, Milwaukee WI. 526 pp.

INTELEX

Is Data An Asset?

- Data is a strategic enterprise asset
- Compared to other assets (cash, property, land, equipment and so on), data is a relatively new asset
- In Clive Humby's words - **"Data is the new oil"**
- In today's competitive world, good data is the key differentiator
- Data needs tender love and care

INTELEX

Why is Data Important?

- Provides statistics as part of an organization’s impact and outcome reporting
- Needed for regulatory reporting
- Insights into trends, behaviours, performance and patterns
- Supports decision-making

Evolution of Data

Evolution of data can be split into three eras

- Before the advent of computers and databases
- After the advent of computers, software systems and electronic storage devices prior to the Internet era
- Internet Era-onwards

INTELEX

Evolution of Data - Before the advent of computers and databases

- No/limited data collection of corporate entities, events, transactions and operations
- Important data were stored in physical (paper) files and registers
- Manual search and retrieval of data from paper files

INTELEX

Evolution of Data - *Advent of computers, software systems & electronic storage devices prior to Internet era*

- Manual entry of data
- Data were collected and stored in electronic files and databases based on business requirements
- Increasing volumes of data input, processed/transformed and stored with the advancement of technology
- Analysis of data using Excel, SQL queries or other programs/procedures

INTELEX

Evolution of Data – *The Internet Era onwards*

- Progress of information technologies, declining cost of disk hardware and availability of cloud storage
- Electronic capture, processing and storage of large volumes of data through multiple channels
- Sourcing and processing massive amounts of data from heterogeneous sources using various software tools and technologies (for example, data warehousing tools and reporting tools)
- Internet of Things (IoT) and Big Data

INTELEX

2: Data Quality and Its Importance

What is Data Quality

- the capability of data to satisfy the stated business and technical requirements of an enterprise
- evaluation of data’s fitness to serve their purpose in a given context

INTELEX

Cost of Poor Data Quality - Business Impacts

- Financial Impacts (for example, increased operating costs and penalties)
- Confidence- and satisfaction-based impacts (for example, customer dissatisfaction)
- Risk and Compliance Impacts (for example, regulatory reporting, non compliance and security breaches)
- Productivity Impacts (for example, increased workload)

David Loshin, 2011, Evaluating the Business Impacts of Poor Data Quality, Knowledge Integrity, http://dataqualitybook.com/kii-content/BusinessImpactsPoorDataQuality.pdf

INTELEX

Cost of Poor Data Quality

A survey performed by The Data Warehouse Institute claims that poor data quality has led to

- lost revenue (54 percent),
- extra costs (72 percent),
- decrease in customer satisfaction (67 percent) (Eckerson).

Eckerson, W. W. 2002. Data quality and the bottom line: Achieving business success through a commitment to high quality data. Renton: The Data Warehousing Institute.

INTELEX

Are the data of high quality? Some points to ponder

- Can the data sources be trusted?
- Are the data relevant?
- Are any data values missing?
- Are data accurate for the context of usage?
- Are the data obsolete or current enough for your analysis purposes?
- Are there any duplicate data records?
- Are the data consistent across systems?
- How timely are the data?
- Are there any security requirements related to the data?

INTELEX

3: Data Quality Dimensions

INTELEX

Data Quality Dimensions

- Data Quality is multidimensional
- Data Quality needs to be broken into measurable characteristics, known as data quality dimensions
- Each data quality dimension captures a measurable aspect of data quality

Mahanti, Rupa. (2019). Data Quality: Dimensions, Measurement, Strategy, Management and Governance. ASQ Quality Press, Milwaukee WI. 526 pp.

INTELEX

Common Data Quality Dimensions

- Completeness
- Accuracy
- Validity
- Uniqueness
- Consistency
- Timeliness

INTELEX

- Completeness - measure of presence or absence of values.
- Accuracy - extent to which data is the true representation of reality, be it characteristics of a real world entity, concept, object, phenomenon or event, which it intends to model.
- Validity - extent of compliance with specific standards (internal/external) or standard data definitions.
- Uniqueness - extent to which an entity is recorded only once and there are no repetitions. Duplication is the inverse of Uniqueness.
- Consistency - extent to which the same data is equivalent across different data tables sources or systems.
- Timeliness - time expectation of availability of data for consumption.

Data Quality (DQ) Dimensions

- Data quality dimensions have been defined and categorized in different ways by various researchers
- There is no universal agreement on data quality dimensions
- Some data quality dimensions are ***objective*** (can be assessed quantitatively), while others are *subjective* (can be assessed qualitatively)
- Examples of Objective DQ dimensions: Completeness, Uniqueness
- Examples of Subjective DQ dimensions: Trustworthiness, Reputation
- You need to define business rules to assess data quality dimensions

Mahanti, Rupa. (2019). Data Quality: Dimensions, Measurement, Strategy, Management and Governance. ASQ Quality Press, Milwaukee WI. 526 pp.

INTELEX

Data Quality Dimensions - Correlations

- There are **correlations** that exist between data quality dimensions.
- For example,
  - The more current the data, the more accurate they are. This is an example of **positive correlation.**
  - Timeliness can have a negative impact on accuracy and completeness. You might have to compromise accuracy and completeness to meet timelines. This is an example of **negative correlation.**

INTELEX

4: Data Quality Management

INTELEX

Data Quality Management

“the management of people, processes, policies, technology, standards and data within an enterprise, with the objective of improving the dimensions of Data Quality that are most important to the organization to achieve the desirable business outcomes”

Mahanti, Rupa. (2019). Data Quality: Dimensions, Measurement, Strategy, Management and Governance. ASQ Quality Press, Milwaukee WI. 526 pp.
Knowledgent, 2017, Building a Successful Data Quality Management Program, Knowledgent Group Inc, Last accessed on January 2, 2017 from https://knowledgent.com/whitepaper/building-successful-data-quality-management-program/

INTELEX

Data Quality Management Approach

- **Proactive -** elimination of data quality problems before they have a chance to appear
- **Reactive - reacting** to data quality issues after they have been introduced

Mahanti, Rupa. (2019). Data Quality: Dimensions, Measurement, Strategy, Management and Governance. ASQ Quality Press, Milwaukee WI. 526 pp.

INTELEX

Proactive Data Quality Management Approach - Examples

- *Train employees to capture data right the first time*
- *Establish consistent standards and data definitions across the enterprise*
- *Establish enterprise data governance*

Mahanti, Rupa. (2019). Data Quality: Dimensions, Measurement, Strategy, Management and Governance. ASQ Quality Press, Milwaukee WI. 526 pp.

INTELEX

Reactive Data Quality Management Approach - Examples

- *Monitoring data quality at different touch points and regular intervals and raising flags in case of data quality issues*

- *Data Cleansing: process of detecting and correcting (or removing) inaccurate duplicate records from a data set or populating missing data values in a data set*

Mahanti, Rupa. (2019). Data Quality: Dimensions, Measurement, Strategy, Management and Governance. ASQ Quality Press, Milwaukee WI. 526 pp.

INTELEX

Hybrid Data Quality Management Approach

- Mix of Proactive and reactive techniques and methods
- *The intent is to be as proactive as possible, but reactive too to manage data quality issues that cannot be prevented. For example, changes in name due to marriage cannot be handled proactively, but the correct date of birth can be captured proactively*

Mahanti, Rupa. (2019). Data Quality: Dimensions, Measurement, Strategy, Management and Governance. ASQ Quality Press, Milwaukee WI. 526 pp.

INTELEX

Six Sigma DMAIC approach to Data Quality Management

- Five phase process to improve and maintain data quality:
  - **Define**
  - **Measure**
  - **Analyse**
  - **Improve**
  - **Control**

INTELEX

- **Define** – Define the problem. Define the high priority data quality issue, the underlying data elements that are causing the issues and the data quality dimensions that would need to be measured and the data quality thresholds.

- **Measure** – Measure the current state of data quality along the data quality dimensions established in the measure phase. Data profiling tools can be used for measurement.

- **Analyse** – Analyse the results of the *Measure* phase to understand the gaps and the underlying causes and propose and assess data quality improvement solution options and choose the optimal solution or solutions.

- **Improve** – Implement the data quality improvement solution. For example, technical solution(s) and/or process change solution(s) depending on the outcome of the Analyse phase.

- **Control** – Sustain data quality improved in the Improve phase; for example monitoring the underlying data elements.

Mahanti, Rupa. (2019). Data Quality: Dimensions, Measurement, Strategy, Management and Governance. ASQ Quality Press, Milwaukee WI. 526 pp.

INTELEX

Data Quality – Some concluding thoughts and tips

•Not all data are equally important. Target only critical data elements (CDEs) for data quality purposes.

•Understand the context of data usage and target the data quality dimensions that need to be taken into account to address the business need.

•Achieving zero data quality issues is a difficult and costly exercise and you do not need 100 % data quality. Engage the stakeholders to establish the minimum acceptable data quality levels for the data set/data elements under consideration.

•Establish enterprise data governance which is rigorous in its definition and enforcement of data standards and policies, clear accountabilities and responsibilities around data – Without data governance it is very difficult to succeed in achieving and maintaining data quality .

INTELEX

This is not the kind of book that you ll read one time and be done with. So scan it quickly the first time through to get an idea of its breadth. Then dig in on one topic of special importance to your work. Finally, use it as a reference to guide your next steps, learn details, and broaden your perspective.
from the foreword by Thomas C. Redman, Ph.D., 'the Data Doc'.

From the Back cover-
Good data is a source of myriad opportunities, while bad data is a tremendous burden. Companies that manage their data effectively are able to achieve a competitive advantage in the marketplace, while bad data, like cancer, can weaken and kill an organization.

In this comprehensive book, Rupa Mahanti provides guidance on the different aspects of data quality with the aim to be able to improve data quality. Specifically, the book addresses:
-Causes of bad data quality, bad data quality impacts, and importance of data quality to justify the case for data quality
-Butterfly effect of data quality
-A detailed description of data quality dimensions and their measurement
-Data quality strategy approach
-Six Sigma - DMAIC approach to data quality
-Data quality management techniques
-Data quality in relation to data initiatives like data migration, MDM, data governance, etc.
-Data quality myths, challenges, and critical success factors

Students, academicians, professionals, and researchers can all use the content in this book to further their knowledge and get guidance on their own specific projects. It balances technical details (for example, SQL statements, relational database components, data quality dimensions measurements) and higher-level qualitative discussions (cost of data quality, data quality strategy, data quality maturity, the case made for data quality, and so on) with case studies, illustrations, and real-world examples throughout.

Mahanti, Rupa. (2019). Data Quality: Dimensions, Measurement, Strategy, Management and Governance. ASQ Quality Press, ISBN:978-0-87389-977-2 Milwaukee WI. 526 pp.

INTELEX

CONTACT

Rupa Mahanti

Business and Information Management Consultant and Author
Associate Editor, *Software Quality Professional*

rupa.mahanti0@gmail.com

https://www.linkedin.com/in/rupa-mahanti-62627915/

INTELEX

Supplemental Slides

•Critical Data Element – A data element that supports critical business functions/processes/enterprise obligations and will result in customer dissatisfaction, pose a compliance risk and or have a direct financial, regulatory or reputational impact if its quality is not up to the mark along one or more data quality dimensions.*

•Data Profiling – Process to capture statistics that help understand the data available in an organization, provide picture of the current state of its data assets and provide some useful characteristics of the underlying data.*

•Data Quality Threshold – The degree of conformance required for the data to be considered of acceptable quality.*

•Data Governance - The exercise of authority, control and shared decision making (planning, monitoring and enforcement) over the management of data assets.**

•Big data is data that contains greater variety arriving in increasing volumes and with ever-higher velocity. This is known as the three Vs.***

Mahanti, Rupa. (2019). Data Quality: Dimensions, Measurement, Strategy, Management and Governance. ASQ Quality Press, Milwaukee WI. 526 pp.
**DAMA Dictionary of Data Management
***Gartner, https://www.oracle.com/au/big-data/guide/what-is-big-data.html

INTELEX

THANK YOU!

INTELEX