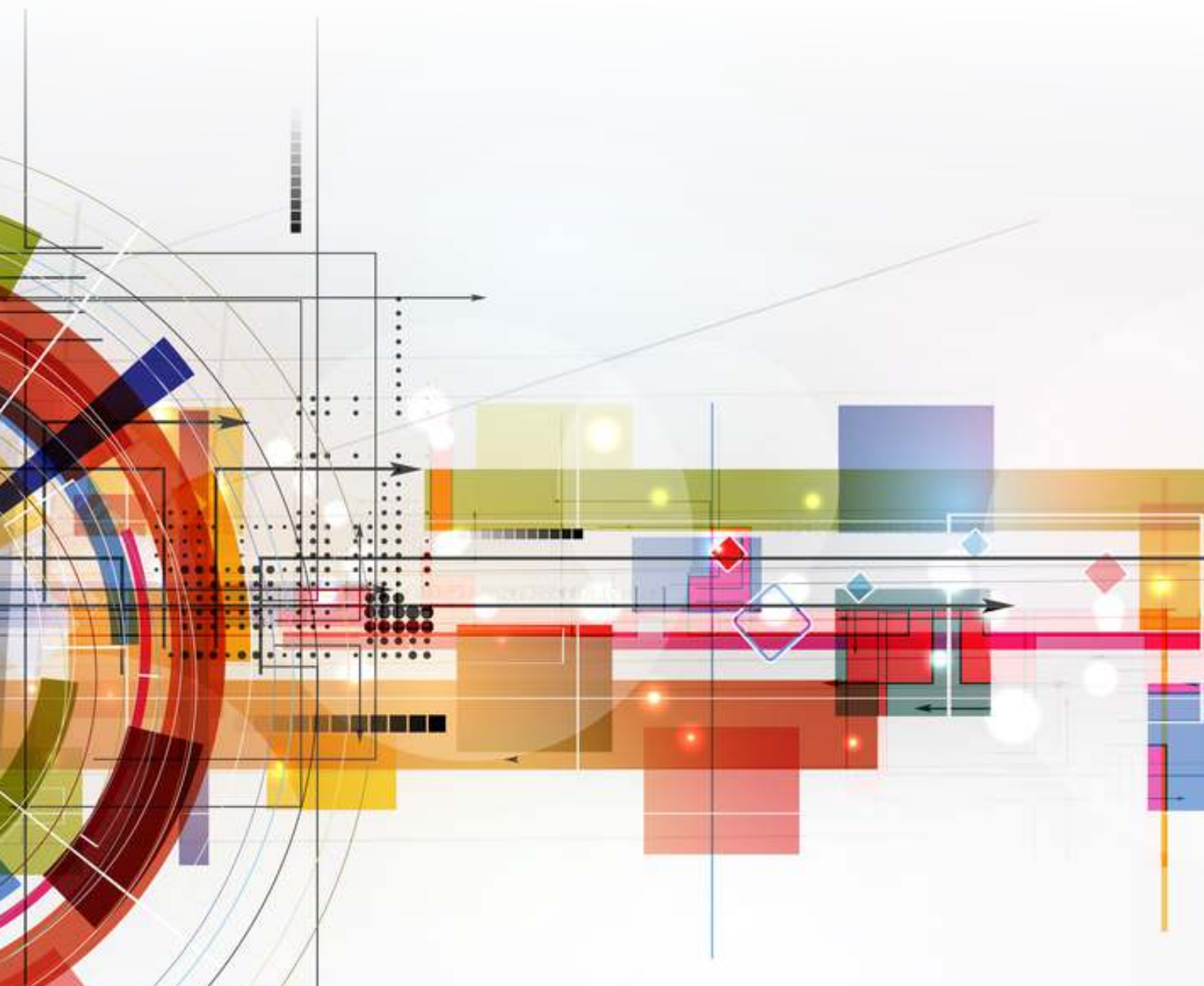
SPAN
GLOBAL SERVICES

Data Cleansing, Quality &Enrichment

Clean up your data act or face the consequences

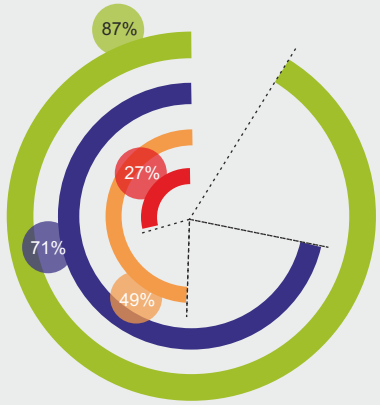


Introduction

Poor data quality is the primary reason for 40% of all business initiatives failing to achieve their targeted benefits (according to Gartner). It has been found that data quality problems cost 10% of the total revenue. The staff of an organization spends 25% of its time in handling customer complaints caused by erratic data, fixing incorrect data, finding missing data, and clarifying data that doesn't make any sense.

BAD DATA – BAD IMPACT!

Respondents in a survey said that poor data quality causes



- 71% - Financial reporting problems
- 49% - Operations problems from bad data
- 87% - Excess costs for business operations
- 27% - Strategic planning problems

Source -

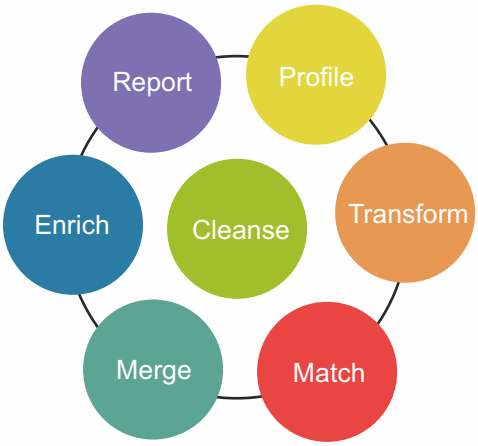
http://www.mytechanalyst.net/2013/10/07/baddata-is-a-social-disease-part-1/

With the evolution of information age, there is enormous amount of data which has been made readily available through the virtual media and sophisticated communications network. There is a widespread conceptual application of internet in supply chains, data mining, knowledge management and many other related concepts. There has been a considerable increase in independent distributed database servers that directly provide online info-retrieval services to end users This has facilitated information manipulation of multi-source data to a great extent. The resulting integration of data from multiple independent sources, results in certain incompatibilities. Most of the users neglect the importance of data in computers but data acts as the real fuel in the information technology engine. Incorrect and futile data should be removed as it results in faulty analysis.

Inaccurate data leads users to make faulty decisions. The impact of data errors is felt by everyone at one time or another, no matter to which function of work the user belongs – management, internet applications, financial applications, marketing or information services. Mostly, poor data quality results in loss of time, money and the effort behind critical business decisions.

What is Data Cleansing?

Data cleansing, also called data scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data. When multiple data sources need to be integrated, e.g., in data warehouses, federated database systems or global web based information systems, the need for data cleansing increases substantially. This is because the sources often contain redundant data in different representations. In order to provide access to accurate and consistent data, consolidation of different data representations and elimination of duplicate information become necessary.



A data cleansing approach should satisfy several requirements:

1. Detect and remove all major errors and inconsistencies both in individual data sources and when integrating multiple sources. The approach should be supported by tools to limit manual inspection and programming effort and be extensible to easily cover additional sources.

2. Data cleansing should not be performed in isolation but should be based on comprehensive metadata. Required mapping functions for data cleansing in accordance with the Meta data should be specified and be reusable for not just other data sources but also for query processing. Especially for data warehouses, a workflow infrastructure should be supported to execute all data transformation steps for multiple sources and large data sets in an efficient manner.

Why are Data Cleansing, Quality and Enrichment important?

According to a Market report, the greatest barriers to B2B lead generation are predominantly the result of bad data quality. With such huge repercussions that the business needs to face due to poor data; Cleansing, Quality and Enrichment gain importance. Most functions in business are data centric from accounting to highly creative department marketing. With such emphasis on good data leading to great business performance it is critical to ensure the data complies with the industry standards.

TRADITIONAL APPROACH	NEW PERSPECTIVE
Data actively collected with user awareness	Most data from machine to machine transactions & passive collection - difficult to notify individuals
Definition of personal data is predetermined and binary	Definition of personal data is contextual and dependent on social norms.
Data collected for specified use	Economic value and innovation come from combining data sets and subsequent uses.
User is the data subject	User can be the data subject, the data controller, and/or data processor
Individual provides legal consent but is not truly engaged.	Individuals engage and understand how data is used and how value is created
Policy framework focuses on minimizing risks to the individual.	Policy focuses on balancing protection with innovation and economic growth.

With so many sources for the collation of data, the probability of making errors like duplication and missing of fields is high. Also with the emergence of cloud-based data collection techniques the reasons and the approaches for data collection have also changed. Now, data collection is not done just with the intention of implementing a particular process but also helps in making strategic business decisions. These decisions determine whether or not a process is worth the effort. From pre-planning to the execution and further to the result; data drives the whole cycle.

Phases in Data Cleansing

In general, data cleansing involves several phases

Data analysis:

In order to detect which kinds of errors and inconsistencies are to be removed, a detailed data analysis is required. In addition to a manual inspection of the data or data samples, analysis programs should be used to gain metadata about the data properties and detect data quality problems.

Data de-duplication:

It is a specialized data compression technique for eliminating duplicate copies of repeating data. This technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. In the de-duplication process, unique chunks of data, or byte patterns, are identified and stored during a process of analysis. As the analysis continues, other chunks are compared to the stored copy and whenever a match occurs, the redundant chunk is replaced with a small reference that points to the stored chunk. Given that the same byte pattern may occur dozens, hundreds, or even thousands of times (the match frequency is dependent on the chunk size), the amount of data that must be stored or transferred.

Data Standardization:

Data standardization is the first step to ensure that your data is ready to be shared across the enterprise. This establishes trustworthiness for the data to be used by other applications in the organization. Ideally, such standardization should be performed during data entry. If, for some reason this is not possible, a comprehensive back-end process is necessary to eliminate any inconsistencies in the data. Data standardization is usually done under name heads like Name, Address, Product, Business, Financial etc.

Data normalization:

It is the process of organizing the fields and tables of any relational database to minimize redundancy. Normalization usually involves splitting of larger tables into smaller (and less redundant) tables and mapping relationships between them. The objective is to isolate data so that additions, deletions, and modifications of a field can be made in just one table and then propagated through the rest of the database using the mappings created.

Quality Check:

This should be performed at every level of Data cleansing. But for obvious reasons, like the source and timing of the data receipt might vary largely, making it difficult to check and assure the quality at the very beginning. Hence an assigned stage of quality check becomes imperative. A typical outline of the timeline at which a quality check process is assigned is illustrated below –

Data generation	• Different sources • Different times of data collection
Different file formats	• Excel files • Text files
Application program functionalities	• Importing data from different files • Filling data gaps • Quality Check • Derived parameters

Benefits of Data Cleansing

There are plenty of financial benefits to cleansing your databases, not only will such activities stop you wasting money on storing old or obsolete data, but the improved accuracy can result in greater return on investment for your marketing activities.

Assists in your target profiling, meaning that your marketing campaigns will be more focused, relevant and subsequently more likely to be successful. Understandably, using obsolete data for your targeting activities will only lead to misdirected campaigns

Storing old information can actually harm your brand image and customer perception. For instance, if you are consistently sending marketing literature to old addresses, customers that may have died or opted out of receiving promotional material, your efforts are actually helping to create a wider negative image, rather than return on investment

- In compliance to the Data Protection Act with regards to the storage of personal information, cleansing activities are necessary to eliminate any chances of violation of the act and subsequent penalties

- Finally, improving the quality of your data results in improved customer insight and adds considerable value to your marketing efforts by improving ROI

These are just some of the reasons to undertake data cleansing, however, as a business, it should be remembered that the process is ongoing. Databases require practically constant management and cleansing to ensure they contain, up to date, accurate and valuable information. In short, improve the quality of the data.

The 5 key issues with data quality

ISSUE #1: Inconsistent Data

Data storage at more than one place results in data inconsistency. Thus on some occasions changes may not be made at all places which results in lack of data integrity.

The average duplication ratio in vendor names is more frequent than that of software product names which is 10:1 than 20:1 in the latter. This is often observed because the consistency across names is difficult to assign. Most software product names are trademark or registered entities.

ISSUE #2: Duplicate or Conflicting Data

Most of the relational databases are a combination of data from multiple sources; they often contain duplicate data. If the database is to be the authoritative system of record supporting IT decisions and processes, it needs to resolve those conflicts and filter out duplicates. The problem and the impact can be huge. Different inventory and asset management tools may insert duplicate data, for many reasons. When the problem scales up, searching and fixing the duplicates/conflicts can be a huge task.

An unbiased research of the industry conducted by BDNA shows that 40 percent of data collected across different sources is repeated.

ISSUE #3: Irrelevant Data

Another issue is data relevance. Removing the irrelevant data can reduce the data footprint significantly. For example, when looking at data for an audit, you do not need to look at files that are individual components of a larger bundle. Ignoring those files helps to eliminate the frequency, so that you can focus on the remaining percent of data that is relevant.

A research done by BDNA shows that 95% of data gathered from various discovery sources is irrelevant.

ISSUE #4: Incomplete Data

After you have addressed issues like inconsistency and duplication, the database needs to be looked in for missing data that you need to make decisions. The source of the data has to be authentic for the data to be of the highest quality. Along with the source, the collection systems also largely determine the quality and completeness of the database. IT systems lack critical information you need to know about all of your assets, such as end-of-life or end-of-support date, licensing/packaging options, current version, etc. These external data points – market data – are essential for many day-to-day processes.

There are various categories of data fields that might be incomplete. The probability of missing data should be assessed prior to the database design. For example, a respondent survey about income of an individual has a very high chance of being left incomplete.

ISSUE #5: Incomplete Data

It's mandatory that some of the data in your entire database is now outdated. There is a continuous inflow of some information with the IT systems in place currently. Most data collected and then collated are more a mundane process than a requirement. With such collection systems, we need to determine a threshold after which the data which is not of any relevant use should be destroyed. This would eliminate the problem of outdated data at large.

25 percent of the Fortune 500 use software that is past its support date.

Features that describe data quality

We had a look at the issues due to which the quality of the databases gets hampered. It is important for us to know the essential features that help us in the assessment of the quality of any large data. The four cornerstones of data of the highest quality are

Intrinsic – The data should be accurate. Accuracy covers most of the aspects of duplication, relevance and authenticity. The sources of the data collated should be reliable and reputable. A clear objective needs to be specified for the data to be acceptable as high quality.

Contextual – The right thing at the right time is the key to the context of a database. Timely and relevant value added to the database ensures the completeness of the data. Also, a large database does not mean a database which has fields that are of no monetary or strategic relevance.

Representational – The database needs to be easy to understand and user friendly. A complicated data only uses a lot of more time for the user to access and make good use of the same.

Accessibility – The security and availability of the data over networks and at the required time of access is essential. Data available at the right time enables right decisions.

Dimension	Characteristics
Intrinsic	Believable, Accurate Objective, Reputable
Contextual	Value-Added, Relevant Timely, Complete Appropriate amount
Representational	Interpretable, Easy to understand Consistent , Concise
Accessibility	Available, Secure

Data Enrichment – How is it different from data quality?

The fourth step in an effective data quality methodology involves creating as complete a picture as possible to support strategic decision-making. Data enrichment can range from something as simple as adding a missing postcode, to augmenting records with demographic or geographic data based on a name or address match. It's a simple process, with the value gained far outweighing the effort required.

Enhance and Enrich examples –

Geographic: such as postcode, county name, longitude and latitude, and political district

Behavioral: including purchases, credit risk and preferred communication channels

Demographic: such as income, marital status, education, age and number of childre

Psychographic: ranging from hobbies and interests to political affiliation

Census: household and community data

While most data quality tools focus on name and address data, the data enrichment processes should support any number of non-name and non-address data types, including dates, telephone numbers, account numbers and e-mail information, to name just a few. The process should also enable full customization of the data quality rules engine, so you can modify existing rules, tweak data types and create your own data types. For example, If you want to parse, standardize and match product names, you simply create a "product name" data type that is specific to the type of data you have.

About Us

Span Global Services offers following Data Solutions to keep your data up-to-date to get optimum results and improve ROI. Translate data into actionable insights with our solutions. We provide the following services:

List Management

Email Lists
List Building

Data Management

Data Cleansing

Data Profiling

Data Verification

Data Appending

Email Appending

Phone Appending

Contact Appending

Social Media Profile Appending

Conclusion

The maintenance of data isn't something that occurs only before you use it for marketing purposes. It encapsulates the processes before use of the data and after use of the data. Processes need to be in place where all changes and updates are fed back into the database to ensure accuracy and consistency. Networks are changing all the time, with new devices and software. And the vendors are constantly changing, with new software versions, patch updates, product names, mergers and acquisitions, and support changes. A large organization may have hundreds of updates each week. The industry keeps changing, issuing new releases, while you're gathering the data. Organizations often spend a lot of time and money in the initial setup of the collection of the data and setting up the system to collate. But then they find that the updates are quiet frequent. Any failure in maintaining and updating would lead to outdated data systems and fields. Due to the various issues that impact data quality, only a very small percentage of the overall data is really clean. Most of the remaining data should be discarded. The clean data should then be enriched with market information to give you the data that really matters.

References

1. Data mining and knowledge discovery handbook, Chapter 2, Data cleansing – A prelude to knowledge discovery; Jonathan.I. Maletic, Kent state university.

2. Data Mining and Knowledge Discovery, 2, 9–37 (1998) © 1998 Kluwer Academic Publishers, Boston

3. Data Cleaning: Problems and Current Approaches; http://dbs.uni-leipzig.de

4. Think you have clean data in your CMDB? Think again! www.bdna.com

5. What does Data Cleansing & Enhancement mean and how can I justify it?

www.celsiusinternational.com

6. Data Quality - A problem and An Approach – Whitepaper, Authors - Javed Beg and Shadab Hussain

7. How Data Cleansing always saves money for Mailing Campaigns - http://www.data-8.co.uk

8. Enterprise Data Quality – Viewpoint by Sivaprakasam S.R.

9. Data Quality Management –Maximize Your CRM Investment Return - www.activeprime.com

10. Trends In Data Quality And Business Process Alignment – A forrester research report, November 2011

11. The Cost of Dirty Data – An educational report

https://www.vha.com/Solutions/Analytics/Pages/DataLYNX.aspx

12. The Cost of Poor Data Quality - How to make better business decisions and positively affect your bottom line © The D&B Companies of Canada Ltd

13. A case for Data Cleansing – Whitepaper - www.imaltd.com

14. Data Quality: A Critical Component of Business Assurance -

http://www.dataconsulting.co.uk/Files/wp_data_quality.pdf

15. Data Quality Strategy: A Step-by-Step Approach - http://www.meritalk.com/uploads_legacy/whitepapers/WP3131_A_DQ_Step_Approach.pdf

16. Essential Data Quality Management - http://dma.org.uk/sites/default/files/tookit_files/wp_essential_data_quality_management.pdf

SPAN
GLOBAL SERVICES

Span Global Services
848 N. Rainbow Blvd.
Suite#5439 Las Vegas, NV 89107

USA: 877- 837-4884
Canada : 877-452-2061
UK : +44 (0) 800 088 5015

info@spanglobalservices.com

Span Global Services is a leading provider of digital marketing and data-driven services. The brand's forte lies in its data intelligence, which holds the largest intellectual mapping available in the industry. As an expert B2B marketing solutions provider, Span Global Services specializes in customized services using the latest business models in online marketing, search marketing, and innovative data strategies. It is the world's only social verified and email verified data provider today. With nearly a decade's expertise in digital marketing, its business intelligence enables companies to utilize the intellectual online marketing strategies along with data insights, market reports, and IT support services. Consulting, Marketing, or Outsourcing solutions — Span Global Services is the most preferred choice.