Tackling The Challenges Of Dirty Data

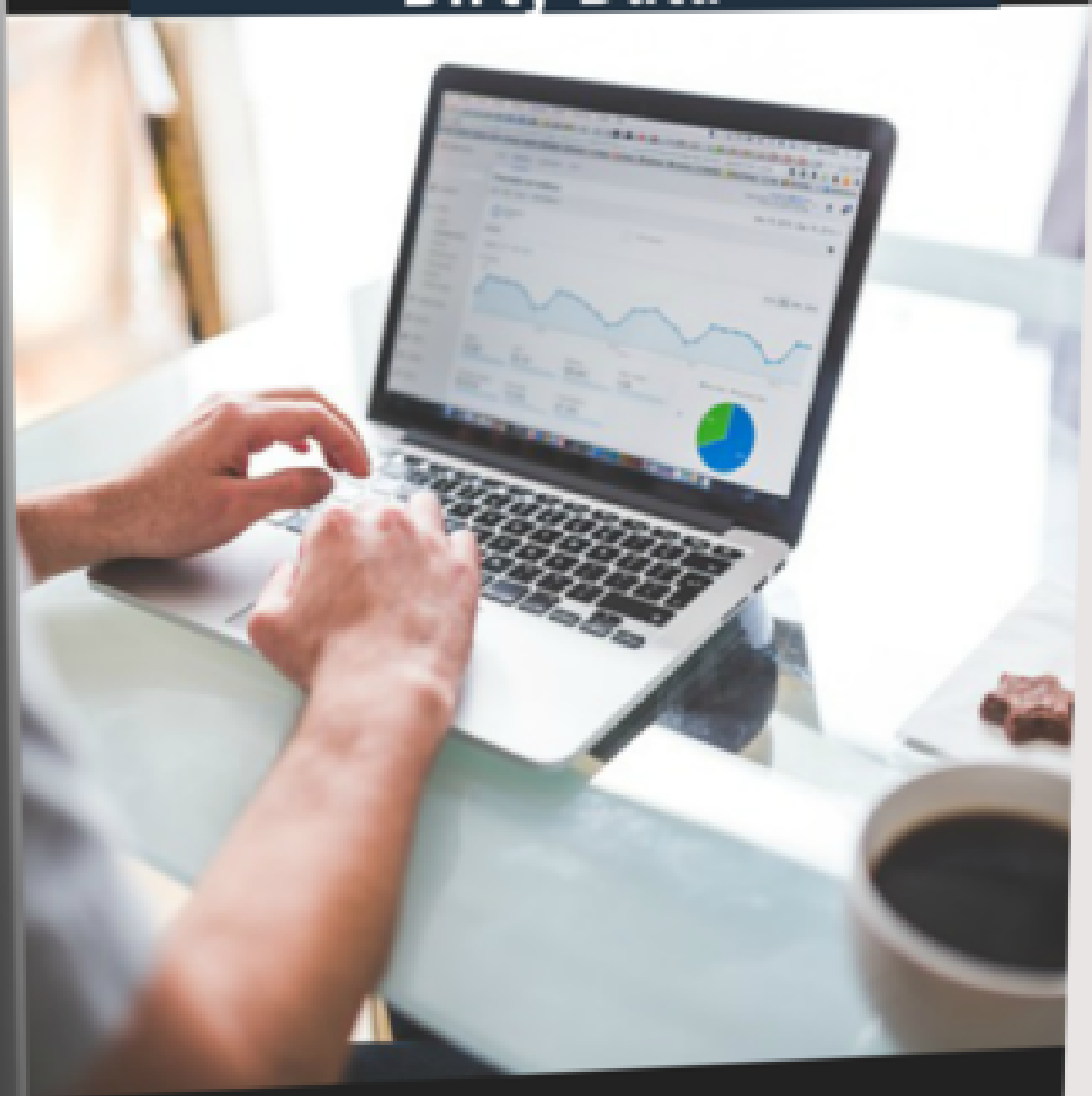

Table Of Contents

What Is Dirty Data?

Here's How To Tackle Your Organization's Dirty Data

How AI Can Clean Data And Save You Man Hours And Money

What is 'Dirty Data'?

CHAPTER ONE

A major concern of CTOs and CEOs of Enterprises that have deployed data analytics is dirty data and its consequences. After all, data quality is one of the top three challenges that Enterprises face in their business intelligence programs.

All of you know that to utilize Big Data, it needs to be viewed within the framework of business context. The bridge between the two is master data management. This kind of program aims to club together data from disparate sources and establish its veracity to ensure data consistency.

Off the starter's block, the first thing businesses usually do is to perform data asset inventories. This helps establish the baseline for relative values, uniqueness and validity of all incoming data. Going forward, these baseline ratings are used to measure all data.

Sounds good, right? Unfortunately, data asset inventories by themselves are simply not enough. Organizations run into hurdles as they grow, and the more the growth, the more the chances of data quality getting compromised.

There are many reasons for poor data. Top of the list is human error. Starting from typos to erroneous values being entered to the duplication of entries; the list is long.

Just below human error is IT architecture challenges. IT relies on multiple hardware and software platforms and solutions. If their mesh up is not done properly, it could cause data problems. What's more, not updating systems as the Enterprise grows could also add to consistency errors.

> *Think of it as a pipeline bringing in the crude oil to the refinery. You need to monitor it for leaks, spillages, corrosion, joint failure, metal fatigue, human error; all of which can led to contamination or leakage at the ingest stage itself.*

Then there is the problem of data decay, which not many businesses factor in.

Here's a simple example: a client has moved house and his address on file is no longer valid. Someone in the team needs to be tasked with the duty of keeping the data current. Multiply such errors by the thousand and suddenly you have a massive problem of bad data on your hands.

Areas that get impacted because of bad data

There are overt and covert consequences of poor data getting into your system. Poor data comes with a cost, both tangible and intangible, and monetary as well as to reputation.

The most obvious sphere of business intelligence to be impacted is strategy, followed by decision making. Any strategy born from poor data has to fail. It also does not need rocket science to understand that poor data means poor decisions.

The impact of poor data can also lead to system inefficiencies and negatively affect productivity. All of which will eventually be a strain on your Enterprise's operational costs.

Here's an example of low trust: The number of cars on the road as on a particular date in a particular geography was erroneously filled in as X value rather than Y. The strategy of an automobile manufacturer to introduce electric cars as replacement in that particular market, based on this wrong input will be flawed.

Another example: Faced with a deadline, a research executive for a detergent company keys in an approx. number under the men-women ratio in a particular suburb, without doing the leg work. What will follow next when the company wants to launch a new product is anybody's guess.

The intangibles are difficult to measure but equally important. Consistent bad or poor data reporting in an organization can affect employee morale. Having to deal with inaccurate data 90% of the time can be a frustrating experience. What is needed eventually is one source of truth.

In this chapter, we will learn how data quality can be preserved, and the role of artificial intelligence in helping preserve large data sets.

One of the first things any organization has to do to stop dirty data from getting into the system is to set up a company culture which is data friendly. Whether yours is a big, medium or small sized Enterprise, it is very important to have a culture to encourage the proper use of data and for its analytics.

But the term 'culture' can be vague and intangible. So here's a quick definition – the Enterprise head honchos need to constantly talk to employees at all levels about the use of data analytics and its benefits. This should also include getting across the importance of accurate data and the harmful effects of dirty data. What's more, roles and tasks have to be assigned to designated members of the team; more so to those who will be responsible for ensuring the consistent accuracy of all incoming data.

Data quality and its importance, and its link to data management cannot be stressed enough. Managements need to constantly emphasize to all team members that in order to make the right data-driven decisions, the first and foremost task is to get the correct data in. That's also because the data gets connected from the master database across a company's CRM, DRM and other such services so data management requires a certain consistency across all. Wrong or poor data can even become dangerous for an Enterprise's survival.

By now, you would have also understood that monitoring incoming data for inconsistencies or other errors is not a one-off thing. It is always a priority.

Once you have done all of this and are confident that the message relating to dirty data and its avoidance has gone down the file and rank, you can go ahead and make your investment on creating the processes, including the software and assigning the people to man them. Yet, do not forget, after this too, you will need to continuously monitor and better the quality if incoming data.

Here are some common roles to ensure master data management

There are 3 ways this can be trifurcated: You can have a 3-tiered structure comprising a data owner, a data steward and a data manager. At first, these may sound like overlapping or similar roles, but they are not.

For example, here's what the data owner does:

- Play the pivotal role for data domains
- Defines data requirements
- Preserves data quality and accessibility
- Decides who in the team gets what kind of access rights and
- Permits data stewards to manage data

So the data owner operates at a macro level in the ecosystem.

On the other hand, the data steward is actually the one who lays down the rules, and the plans, and then coordinates data delivery. He is the operations guy.

Last in this change is the data manager. This guy operates at a micro level in the Enterprise. He normally coordinates between the data owner, and the technical part of implementing the plans.

Now that you have the systems, procedures and manpower in place, what next? Remember, a fruitful data quality and management project requires a holistic approach.

This is where the 'How' part comes in – how to go about ensuring consistent good data?

To determine the quality of data, here are some aspects to look out for:

Accuracy, completeness, adherence to standards and duplication. A combination of IT software, hardware and human resources will take care of this. Your designation team with the given infrastructure first needs to identify all the problems areas from where bad data is likely to come in. Remember, all this effort is towards establishing a single source of truth.

Thereon, your Enterprise will then have to develop a data quality program, and with the help of a data steward, apply the business processes that ensure all future data collection/ use meet regulatory frameworks and eventually adds value to the business at hand.

The correct method of matching high data quality with technology is to integrate the different stages of the data quality cycle into operation procedures, and tie them in with the individual roles.

Use of AI in manipulating large data sets

In an earlier post, we had written how with the entry of AI, data stewards could now use data cleansing and augmentation solutions based on machine learning (ML).

ML and deep learning allow the analysis of the collected data, and make estimates to learn and change as per the precision of the estimates. As more information is analyzed, so also the estimated progress.

While identifying where your data is lacking or erroneous, large data sets always present a problem. How do humans track say a million data points? And, in real time? But with ML getting into the mix, that hurdle, too, can be surmounted. AI can be used to detect anomalies in data sets by being “trained” to continuously track and evaluate data, even as the data is being processed.

What is even more important that an ML solution can detect and deal with data integrity issues at the very start of data processing, and quickly convert such vast volumes of data into dependable information.

In conclusion: Tracking, analyzing and correcting/updating incoming data will eventually help an Enterprise in taking well-informed business decisions, providing a single source of truth and eventually increased productivity.

References: https://orangematter.solarwinds.com/2019/01/24/how-to-clean-dirty-data-the-life-of-a-data-janitor/

https://www.dnb.com/perspectives/master-data/6-key-responsibilities-of-data-stewards.html

https://www.newgenapps.com/blog/data-cleaning-with-ai

https://searchdatamanagement.techtarget.com/definition/data-quality

https://www.information-management.com/news/poor-data-quality-causing-majority-of-artificial-intelligence-projects-to-s
tall

Image by Gerd Altmann from Pixabay

https://orangematter.solarwinds.com/2019/01/24/how-to-clean-dirty-data-the-life-of-a-data-janitor/

https://www.dnb.com/perspectives/master-data/6-key-responsibilities-of-data-stewards.html

https://www.newgenapps.com/blog/data-cleaning-with-ai

https://searchdatamanagement.techtarget.com/definition/data-quality

Image by S. Hermann & F. Richter from Pixabay

Dirty data is the bane of the analytics industry. Almost everybody or organization that deals with data has had to deal with some degree of unreliability in their numbers.

There are studies out there that tell us that Enterprises spend as little as 20% of their time analyzing data, while the bulk of the time is spent in cleaning or prepping it.

Needless to say, poor data leads to poor interpretation of it. The result – assessments based on such data become faulty, and often leads to the missing of enterprise goals, or increased operational cost, or even customer dissatisfaction.

“Poor” data is a very generic way to describe data inconsistencies. It can be compartmentalized into the following:

Duplicate data: One event is entered twice or more over in the dataset

Missing data: Values are missing from some of the fields

Invalid data: Data entered is old or incorrect

Bad data: Means all kinds of formatting problems, including spelling errors

Most data errors creep in because of humans. Ensuring data quality is an irksome process, requiring monotonous and repetitive actions, which if not done as per the book, leads to bad data slipping in.

Ironically, you need to do an analysis at the start itself to understand the type of irregularities and errors that are likely to creep in, and that need to be removed during the collection and recording of all data. Best practices have to be deployed vis-à-vis data quality and governance at this point in the chain.

It's a two-step action plan. The first step in the data analytics process is to identify bad data. The second involves the fix; meaning, corrective action such as deletion, and replacing the deleted data with another dataset.

Before the advent of artificial intelligence (AI) and its subset machine learning (ML), data analytics companies had no choice but to use the traditional data cleansing solutions to do the job. The problem, though, with such solutions is that they are more or less, rule-based, and are not very scalable. They fall by the wayside, exhausted and unable to cope with the large volumes of data pouring in.

But the entry of AI now means data stewards can use data cleansing and augmentation solutions based on machine learning.

Machine learning and deep learning allows the analysis of the collected data, make estimates, to learn and change as per the precision of the estimates. As more information is analyzed, so also the estimates progress.

So how does it really work?

Since data flows in from a disparate number of sources, any program using ML requires it to get the data into a stable arrangement to simplify it, and ensure same patterns across all points of data collection.

Depending on the number of data sources, the level of diversity, etc, steps may have to be initiated to transform the data. At this point itself, the suitability of the transformation activity and the definitions must be analyzed.

Once this is done, the bad data must be substituted with the good data in the primary source. This is a very important step as it means all data across the enterprise is refreshed, permeating throughout all the divisions, removing any need to removals in the future.

> *Human error is found to be the main reason in critical areas of data collection so any AI based model uses ML to replace humans in identifying bad data and refreshing the models as and when needed.*

ML is also used to establish the "errors" a data analytics models was likely to produce.

Obviously, the more information that is offered, the better it gets. Which means contrary to manual cleansing systems, the ML based algorithm gets better with scale. Thus, as the ML based software gets more and more "smarter" because of deep learning over time, less time is spent on the cleaning of data even as it is flowing in, speeding up the entire data delivery process.

Replacing humans with "machines" also guarantees:

- Clean data
- Standardization of data
- Reduction of coding hours and time spent correcting faulty data at source
- Allows customers to integrate their 3rd party apps with such cleansers

Such ML-based programs use the Cloud and on-premise delivery models, can provide customizable data solutions – which means any enterprise across verticals like marketing or healthcare can deploy it – and offer better metadata management abilities to provide better data governance.

Image by Gerd Altmann from Pixabay

References:

1.https://www.newgenapps.com/blog/data-cleaning-with-ai

2. https://channels.theinnovationenterprise.com/articles/is-artificial-intelligence-necessary-for-effective-data-cleanups

Thank You For Downloading This E-Book

We hope you enjoyed reading it

You may get in touch with us at
marketing@expressanalytics.net