OCTOPAI

Make Better Data-Based Business Decisions in 2020

2020

Modern business thrives on a steady diet of data. Decisions large and small rely on proper analysis of the right data.

But just as a diner at a restaurant can fall sick from improperly cooked food, businesses can suffer significant harm from decisions based on incorrect, incomplete, or poorly analyzed data.

The Effects of Bad Data

The history of business is littered with examples of decisions gone awry as a result of poor-quality data and analysis:

- IBM estimated that bad business decisions resulting from data problems cost the U.S. economy $3.1 trillion per year.[1]

- Experian Data found that 88% of U.S. businesses suffer losses as a result of poor data-driven decisions, to the tune of an average of 12% of total revenue.[2]

- Gartner estimates that 40% of businesses fail to achieve their goals because of bad data. The same research found that 40% of a typical company's data is incomplete, incorrect, or flawed in other ways, and that bad data results in a 20% labor productivity decrease.[3]

Just having access to data is not enough. There has to be a process that identifies the quality of data you have as well as what it means in the greater context.

- **Hawaiian Airlines:** In a bad-data double-whammy, in the space of one week in 2019 Hawaiian Airlines charged customers in dollars when they were attempting to exchange frequent-flyer miles for award flights (in one case charging a customer's credit card $674,000), and charged other customers zero miles for award flights (and then canceled the errant tickets, much to the customers' dismay).

- **Facebook:** In 2016, advertisers and publishers on Facebook sued the social network because of Facebook's inflated average viewership metrics, which resulted in higher prices charged to the advertisers. According to the suit, the viewership numbers were overstated by up to 900%. Whether the figures were deliberately and systematically inflated or simply the result of poor analytical practices, the cost to Facebook's reputation among advertisers was significant.

- **"Jobs" Search:** A group of academic researchers attempting to use social media to predict the U.S. unemployment rate (by searching for employment-related words in postings) found that their results were skewed by one event - the death of Apple's Steve Jobs, which caused a spike in the word "Jobs" on social media.

OCTOPAI

Other cases where bad data or poor analytics caused embarrassment are also common, for example:

- The Target department store chain infamously used its analysis of buying patterns to determine that a specific customer was pregnant, and sent coupons for baby items to her home. However, she was a teenaged girl who had not yet told her parents she was pregnant, and the coupons made a stressful situation even worse. The analysis was correct, but the assumption that pregnancy is a happy event for everyone was short-sighted.

- Social media giant Pinterest sent emails to hundreds of people congratulating them on their upcoming weddings and advertising custom wedding invitations. The problem was that many of the recipients were already married or had no plans to get hitched. Unlike the Target example, the analysis here was poor.

These are the issues that keep corporate leaders up at night, worrying that their decisions are being driven by data and analysis that are fundamentally flawed.

The Many Potential Problems with Data

How do companies get themselves into so much data trouble?

Sometimes, as in the case of Target, and Pinterest, the trouble starts with bad assumptions and inappropriate analytical methods. All too often, as in the case of Enron and its not-so-independent auditor Arthur Andersen & Co., the culprit is deliberate fraud and deception.

Most often, however, the problem is bad data: haphazard collection and storage practices, "dirty" data values, poor data management and governance, and too many manual processes. Let's look at each of these problem areas in detail.

Collection and storage: Especially in large organizations, data is collected in different ways in different parts of the organization and stored in multiple independent data "silos" that are not related to one another. This makes it difficult to find the right data for a given analysis task, and reduces the trust in the data

because of the chance that important data is missing from the analysis.

Dirty data: Too often, organizations assume that their data is "clean"—that is, every required field is populated in every record, and every field has the right data in it (names in name fields, phone numbers in phone number fields, and so on). This assumption is almost always false, especially if there are few or no controls around data entry or collection.

Data management and governance: Organizations that have inadequate or nonexistent policies and procedures around how data is collected, stored, protected, shared, used, and disposed soon find themselves with a disorganized, dysfunctional data landscape. Worse, such organizations can end up on the wrong end of regulatory actions related to data privacy and protection.

OCTOPAI

Taking Control of Your Data Through Metadata Management Automation in 2020

Manual processes: Too many companies rely on tedious, labor-intensive processes to locate, organize, characterize, and scrub their data. By some estimates, over half of knowledge workers' time is spent finding, organizing, and cleaning data to a point where they can perform trustworthy analysis. This is non-value-added time that they could better spend doing more, or deeper, analysis—if they only had data they could trust.

The data situation in many organizations is unsustainable. Companies that can't trust their data cannot make the critical, timely tactical and strategic decisions needed to stay on top of their market and ahead of their competitors.

Happily, there are strategies and tools available to keep you from falling into the bad data trap. All of them are based on a foundation of metadata management.

Metadata, as you may know, is information about the data. Traditional metadata—the type that might be captured in a "data dictionary"—is limited to technical information about the fields in each database table, such as field name, label, data type, format, and field length.

In the context of large organizations and their complex data landscapes, metadata includes much more information, both technical and non-technical. For purposes of this discussion, metadata also includes:

- The source of the data

- The format (such as plain text, comma-separated values, Microsoft Excel, Microsoft SQL Server, and Oracle)

- The status (active for production use, archived, decommissioned, and so on)

- The sensitivity (public, company confidential, contains personally identifiable information, contains customer credit card numbers, and so on)

- Who owns and is responsible for the data (usually a department or team)

- Where and how the data is stored (encrypted or not)

- Transformations that were applied (format changes, extract-transfer-load processes, calculations, joins, and more)

3 Ways to Prevent Data Issues from Driving Bad Decisions in 2020

Metadata discovery and management are the keys to taking control of your data landscape.

With that in mind, here are three important actions you can take around metadata:

Set up a business glossary: Not to be confused with a data dictionary, a business glossary is a listing of standard company-specific business terms and their definitions. This is important for data management because it helps you align disparate data assets so that terms used in one mean the same thing as the same terms in any other. When all data assets adhere to the enterprise's business glossary, it's easier to locate and combine like data from multiple data sources in the same report or dashboard without fear of comparing apples and oranges. Building a business glossary, is an enormous project - but it doesn't have to be. The right automated metadata management tool can generate your business glossary, well, automatically.

Stop manual data mapping: Data mapping—the task of describing how each component (such as a column in a table) in one data object corresponds to a component in another data object—is a mind-numbingly tedious exercise if done by manually examining each database table, spreadsheet, and flat file and trying to connect the dots among them. When transformations need to be done on one or more source objects so that they all agree with the target object, the exercise becomes highly error-prone as well. To make things worse, the task must be repeated whenever new data sources are brought into the picture, such as in mergers and acquisitions. All of this is made much easier by the application of automated data discovery tools.

Deploy robust data lineage capabilities: Data lineage is the ability to trace a data item—for instance, a value in a report—back to its source (or sources). Without an automated data lineage capability, trying to track down the source of an error or discrepancy in a report can be a time-consuming nightmare. With proper graphical data lineage tools, however, it's possible to cut that effort down to minutes instead of days. This gives IT departments and BI professionals several important benefits, including:

- Quick tracing and correction of analysis and reporting errors

- The ability to show auditors that sensitive customer data is kept protected through every step of its journey from source to report or other use

- Rapid assessment of whether an item in a report incorporates all relevant data sources.

OCTOPAI

The Key to Success in 2020: Automation

These actions, along with a robust data governance program, can help get your data environment under control and ready for you to unlock the value within it. And the key to effective implementation of these strategies is the use of automated tools.

It will not surprise you to learn that all of the above recommendations are much easier to implement when they are automated. When automated tools are applied to these tasks, they are completed much more quickly, with much of the "grunt" work done for you. The cases where automated tools are unable to resolve conflicts and ambiguities are few in number, and therefore much easier to deal with.

Furthermore, these automated tasks are easily repeated to keep your metadata up to date when new data sources are added, old ones are retired, and existing ones are modified. Automation reduces errors in the tasks and frees the IT and BI staff for more productive, value-added pursuits.

Bottom Line: Drive the Business Forward in 2020

Getting a handle on your data environment by implementing automated tools for data lineage, data discovery, business glossaries, and data mapping carries with it immense benefits for any enterprise:

- Staff spend less time searching for data (and sometimes failing to find it) across multiple systems, applications, databases, and file servers.

- The organization realizes greater process efficiency, because less effort is spent on dealing with data issues and more can be dedicated to providing value to the customers.

- Reports, dashboards, and customer-facing communications have errors less frequently, and the errors that do occur are easier to track down and correct.

- Because of the reduced error rate, reports are more trustworthy and can be used with greater confidence to inform important business decisions.

- Compliance with various data privacy laws and regulations is easier to demonstrate to auditors and regulators.

Bad data, bad data management, and bad analysis can lead to corporate embarrassment, regulatory actions, and poor business decisions—decisions that can cost businesses money, customers, and reputation, and can even put companies out of business. How do you avoid these issues? Step 1: Make sure you have good data.

As you're reading this, your organization is acquiring ever more data at an ever-faster rate. More data sources are coming online all the time. The problems caused by bad data are not going away and will only get worse for those businesses that fail to get their data under control. Don't fall victim to your own poor data practices—leverage automated tools to implement the recommendations outlined here.

OCTOPAI

Endnotes

1   Thomas C. Redman, "Bad Data Costs the U.S. $3 Trillion Per Year", HBR, September 22, 2016, https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year

2   Matthew Zajechowski, "The Lessons We can Learn from Bad Data Mistakes Made Throughout History", The Smart Data Collective, May 25, 2017, https://www.smartdatacollective.com/lessons-can-learn-bad-data-mistakes-made-throughout-history/

3   Ted Friedman, Michael Smith, "Measuring the Business Value of Data Quality", Gartner, October 10, 2011, https://www.data.com/export/sites/data/common/assets/pdf/DS_Gartner.pdf