Data – The Fundamentals are Broken

Abstract

The world has transitioned from an industrial to an information age, where we have become extremely dependent on information along with the data on which it stands. With the rapid growth of technology, the world continues to move farther away from "hands-on" observation, operations, and management of the real world to utilizing the *data representation of the real world*. Our dependency on data has grown exponentially, while our ability to identify, understand, manage, and utilize data becomes more challenging over time. Despite all the thousands of tools and technologies marketed as "fixes" for the data issues, our data challenges continue to grow. Our growing data issues are only the symptom of the underlying problem – The fundamentals of data are broken.

In this report we have identified the broken fundamentals of data and five basic steps organizations can take to address these broken fundamentals. These include: establishing the business to data connection; using a business blueprint; creating a data oversight framework; establishing an enterprise data construction practice; building the data asset management infrastructure; and, standing-up a data asset management practice focused on enterprise foundational data. In the end, if organizations do not fix the fundamentals, they will never be able to effectively clean, identify, integrate, manage, and utilize their data assets for even basic operations, let alone take advantage of the full power of those assets for true business intelligence, risk avoidance, predictive analytics, artificial intelligence, and data monetization. Information is truly powerful, but only if the data is right.

Author:
Noreen Kendle
Principal Consulting Analyst
Noreen@techvisionresearch.com

Table of Contents

Executive Summary & Key Advice

With the world transitioned from an industrial to an information age we have become extremely dependent on information and the data on which it stands. Data is a representation of the real world–its things events, and relationships. With the rapid growth of technology, the world continues to move farther away from "hands-on" observation, operations, and management of the real world to utilizing the *data representation of the real world.* Our dependency on data has grown exponentially, while our ability to identify, understand, manage, and utilize data becomes more challenging over time. Many of us who are close to the data have a keen awareness of the growing data issues. Despite all the thousands of tools and technologies marketed as "fixes" for the data issues, our data challenges continue to grow.

Our fundamental data problems are not technology problems, yet Information Technology (IT) departments face a growing pressure to implement the latest technology trends, specifically outsourcing, Agile, Big Data, and Cloud. All of these trends place an organization at risk of losing more control of its data, burying itself deeper in broken data, and falling further behind in addressing its data challenges. All of this adds to the dysfunctional cycle of data chaos.

At TechVision, we have come to the conclusion that today's data challenges are a symptom of the problem, not the problem itself. Underlying these data challenges are fundamental principles that are broken. Fixing these fundamentals is critical to getting the data right, managing it as an important business asset, and thus breaking the dysfunctional cycle of data chaos.

As discussed in this report, some of the primary data challenges facing organizations today include: data quality; redundant data; inaccurate, inadequate or missing data definitions; the business data disconnect; and, regulatory compliance data challenges. Nearly all organizations experience these challenges as they spend ridiculous amounts of time hunting for missing data, correcting inaccurate data, creating "workarounds," pasting data together, and reconciling conflicting data. This results in tremendous unaccounted loss in productivity. Rather than fixing the problem most organizations have fallen into the trap of believing these activities are a normal part of doing business. *This is insane*.

Although the concept of data as a business asset is spoken about and may be included in the strategy, data being treated as a business asset is not a reality for most organizations. Nearly all of IT funding, resources, and focused effort have gone into technology, with literally no attention to the actual data itself. Technology advances far surpass the limitations of our data. As each "new" data technology emerges, hits its peak of inflated expectations, and then sinks into disillusionment more continue to emerge. Now that Big Data has passed through its hype cycle, many organizations have come to realize that with big data they face the same challenges of small data but on a much bigger scale–data quality

issues, integration challenges, and insufficient data governance and management. All of the great data technology advancements will continue to fall short of expectations mostly due to the reality of our data–redundant, disparate, inaccurate, missing, misplaced, undefined, mis-defined, dirty, old, reused, misused, etc.

Defining our data challenges and addressing them is essential to gaining an understanding of what caused them in the first place. We must understand the root cause of our data problems so we can get our data right and mitigate this data chaos. We see eight primary causes of these data challenges. These include causes such as the computerization of data; technology driven data; system development practices; the tools technology myth; and, unmanaged data aging. Understanding these challenges and their root cause is the first step to recover. The next step is fixing the broken fundamentals of data. This is a complex process with many moving parts and it potentially touches every facet of the organization.

In this report we outline five core broken fundamentals of data. At the heart of these broken fundamentals is the persistent disconnect between the business and its data. In most organizations data is closely associated with the software application: a usage view of the data. What data represents and its usage are not the same thing. Gaining a holistic view of the business and its data gives the organization a much-needed business perspective of data–bridging the gap between business and its data in a consistent manner. This is the key to turning data into a real business asset. Similar to a technology infrastructure, a data infrastructure is necessary to effectively support the data assets in a consistent manner. The data infrastructure is architected and built based on the holistic view of the business organization that the data represents. The gap between the business and its data is a big issue in most organizations today making this a core reason the fundamentals data are broken.

There are five basic steps organizations can take to address these broken fundamentals. As discussed in this report, these include: establishing the business to data connection using a business blueprint; creating a data oversight framework; establishing an enterprise data construction practice; building the data asset management infrastructure; and, standing-up a data asset management practice focused on enterprise foundational data. In the end, if organizations do not fix the fundamentals, they will never be able to effectively clean, identify, integrate, manage, and utilize their data assets for even basic operations, let alone take advantage of the full power of those assets for true business intelligence, risk avoidance, predictive analytics, artificial intelligence, and data monetization. Information is truly powerful, but only if the data is right.

Introduction

With the world transitioned from an industrial to an information age we have become extremely dependent on information and the data on which it stands. The information age did not mark the beginning of information or data, as data has existed since the beginning of time. Data is a representation of the real world—its things, events, and relationships. The representation of the real world business organization (i.e. the data) is used in nearly every area of an organization, from operations and management of the businesses to its recording history, analyzing the past, making decisions, and predicting the future.

With the rapid growth of technology (e.g., computerization, telecommunications, internet, etc.), the world continues to move farther away from "hands on" observation, operations, and management of the real world to utilizing the data representation of the real world.

Most business professionals trust the data that represents a person, place, or thing without any actual observation or interaction of the real world that the data represents. Our dependency on data has grown exponentially, while our ability to identify, understand, manage, and utilize data becomes increasingly more challenging over time. With the increasing volumes of undefined and unmanaged data, the state of our data is deteriorating. The poor quality, inadequate, and erroneous data (e.g. broken data) is often unrecognized, underestimated, and unmeasured. When we use these broken data to analyze, decide, and predict, most of us remain ignorant to the data's trustworthiness because we assume the data to be correct. The validity of the data is rarely questioned. The resulting effects on business success are unknown, leading to blind ignorance. Unidentified, misunderstood, and untrusted data also affect an organization's ability to secure and protect its data, putting the organization at significant risk—often unknowingly.

> Our dependency on data has grown exponentially, while our ability to identify, understand, manage and utilize data becomes increasingly more challenging over time.

Many of us who are close to the data have a keen awareness to the growing data issues. In spite of all the improvements in technology and the thousands of tools and technologies marketed as "fixes" for the data issues, our data challenges continue to grow. These tools and technologies include object technologies, data quality tools, ERPs, and the Data Warehouse and Star Schemas to EAI, EII, SOA, Master Data, Data Governance, and now Big Data tools. It would be wonderful to simply purchase a tool or technology and have the data challenges disappear, but no tool can remedy the fundamental data problems. The tools and technologies have not failed; they work as intended. Our fundamental data problems

are not technology problems. Rather, our fundamental understanding and approach to data has fallen short.

In today's information economy, organizations can no longer afford to continue the dysfunctional cycle of data chaos where valuable resources are spent looking for missing data, correcting inaccurate data, creating "workarounds," and reconciling conflicting data. This occurs over and over again because organizations are not addressing the underlying cause. To make matters worse, when these organizations use tools and technologies to "address their data issues," the cycle of data chaos expands.

Fundamentally, our data problems are not tool or technology problems. Information Technology (IT) departments face a growing pressure to implement the latest technology trends, specifically outsourcing, Agile, Big Data, and Cloud. All of these trends place the organization at risk of losing more control of its data, burying itself deeper in broken data, and falling further behind in addressing its data challenges. All of this adds to the dysfunctional cycle of data chaos.

> Fundamentally, our data problems are not tool or technology problems.

The traditional approach to data is not working. Something must change. As Albert Einstein said, "We cannot solve our problems with the same thinking we used when we created them." Now more than ever, we need to break this cycle. It is time to step back and take a different look at data.

After many years of dealing with corporate data challenges and experiencing the continual failure of "silver bullet" technology solutions, at TechVision we have come to the conclusion that today's data challenges are a symptom of the problem, not the problem itself. Addressing symptoms never fixes the problem and therefore we look past the symptoms to derive the root cause: underlying our data challenges are fundamental principles that are broken or missing. Our focus needs to shift from the data symptoms to the real problem— the broken fundamentals of data. Fixing the fundamentals is critical to getting the data right, managing it as an important business asset, and thus breaking the dysfunctional cycle of data chaos.

Problem Statement
The Organization's Data Challenges

Data problems are one of the most underestimated challenges organizations face today. Data within any organization is typically disjointed, untrustworthy, redundant, large, and complex. The sheer volume, lack of quality, definition, and mismanagement of data is crippling organizations. Any one of these data problems can greatly hinder an organization's effectiveness, growth, and profitability—perhaps eventually leading to that organization's demise.

Since data is the raw material used to produce information, data is one of the most valuable resources to organizations today. The cost of computer storage and processing has continued to drop as computer storage capabilities continue to increase; this both enables and enhances the growth and importance of data.

With anything of value, there is continual increasing demand. The evolution of analytics requires more data detail. As the demand for data increases, so does its growth rate; fueling a data management nightmare within most organizations, whether it's recognized or not. With data's continually increasing growth rate, and our dependency on the data, any data challenge we face today will pale in comparison to what we will face tomorrow.

> As the demand for data increases, so does its growth rate; fueling a data management nightmare within most organizations, whether it's recognized or not.

Data is a business asset, and for many organizations it is one of their most important assets. When data is compromised, the organization is compromised: customers, stakeholders, employees, and regulators all hold organizations accountable for the accuracy and protection of their data. When inaccurate or incomplete data is used in financial reporting, compliance, transactions, or critical decision-making, an organization can be brought to its knees and possibly even to its demise. When the organization doesn't protect sensitive financial, customer, or employee data, risk increases for theft and fraud. Compromise the data and you compromise the organization it represents.

Nearly all organizations are data challenged: spending ridiculous amounts of time hunting for missing data, correcting inaccurate data, creating "workarounds," pasting data together, and reconciling conflicting data. This results in tremendous unaccounted loss in productivity. Rather than fixing the problem, most organizations have fallen into the trap of believing these activities are a normal part of doing business. *This is insane.*

At TechVision Research we see the following as the primary data challenges facing

organizations today:
- The Data Quality Challenge
- Unknown Inaccuracy Data
- Redundant Data
- Data-Information Overload
- Inaccurate, Inadequate, or Missing Data Definition
- Undermanaged or Unmanaged Data
- Business Data Disconnect
- Regulatory Compliance Data Challenge
- Acceptance of Data Problems as Normal

The Data Quality Challenges

Data quality issues can harm every area of an organization. Data quality refers to the degree of data's accuracy, completeness, relevancy, timeliness, and trustworthiness. Studies regarding data quality have produced disturbing statistics, all of which reveal the costly effects that low-quality data have on an organization. But the statistics do not reflect the true costs that result from the time organizations spend repeatedly correcting inaccurate data, hunting for missing data, and researching data's accuracy. Each one of these activities ultimately detracts from the bottom line.
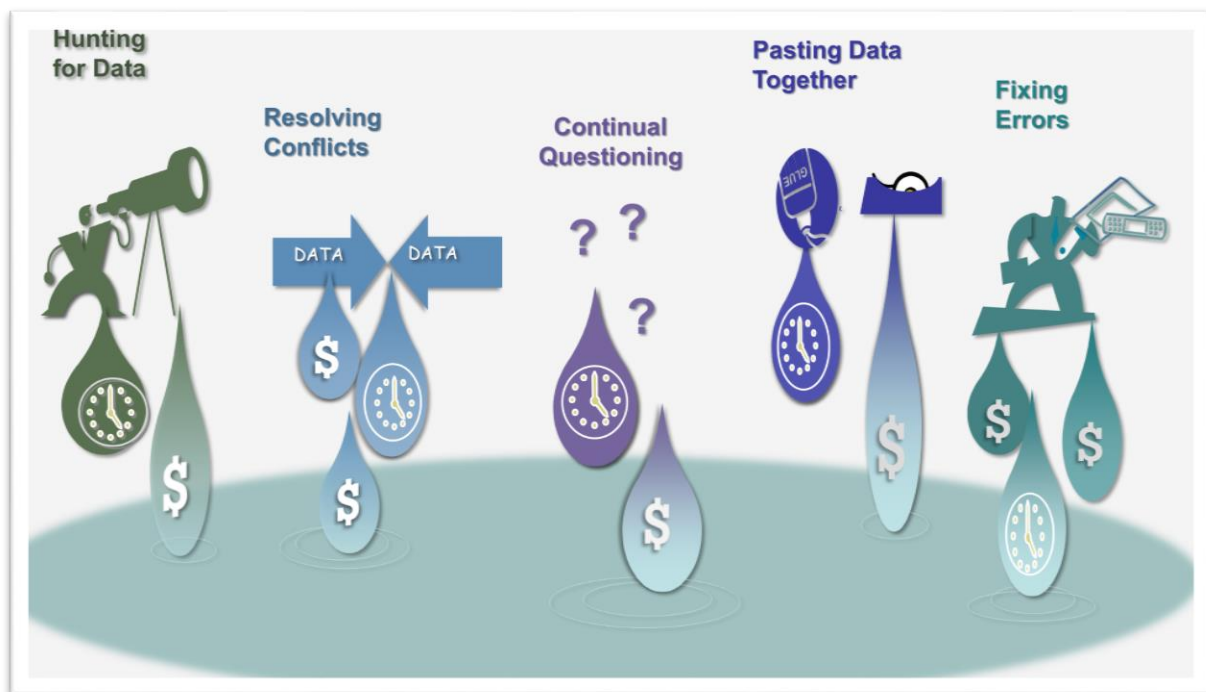
Figure 1: Data Quality Issues Time and Money Drain

Most information workers spend their time verifying the accuracy and quality of the data. Accuracy of data is especially important in making critical operational and strategic decisions that are ultimately fundamental to the organization's survival. Business organizations without an accurate picture of their inventory, profits, losses, or customers will inevitably make poor decisions. Business decisions made with inaccurate data waste time, money, and resources and will eventually lead to business failure.

We see many organizations placing responsibility for data quality on technology, where data quality tools are used to automate identifying data quality issues so the incorrect data can be "fixed" or adjusted. However, as discussed above, poor-quality data is most often a symptom of underlying problems. When organizations treat poor-quality data as the problem, they do not address the underlying problem. The end result is that the underlying problem continues to generate poor-quality data. Time is also a critical factor, since the longer incorrect data remains in a system, the greater the impact. Unfortunately, this scenario is all too familiar in most organizations and results in significant data quality issues.

Unless the organization addresses the underlying cause that produces the bad data, a quick-fix will merely mask the data errors flowing through the organization. To complicate matters, as low-quality data moves through the organization, the data quality further deteriorates because data is often used to make more data.

The 1-10-100 rule of quality describes the increasing cost of a defect as it flows through its lifecycle and is left unfixed. The rule states that for every dollar spent on prevention of quality issues, $10 will be saved in tracking and correcting a mistake. If a quality issue goes undetected until a customer discovers it, the cost of rectifying the mistake will be 100 times what it would have been to prevent it. The 1-10-100 rule applies to data and, in terms of data quality, this means that the cost of undetected data quality issues is 100x the cost of preventing the data quality issue in the first place!

> Organizations are not only challenged by their incorrect data, but even more with their ability to recognize it.
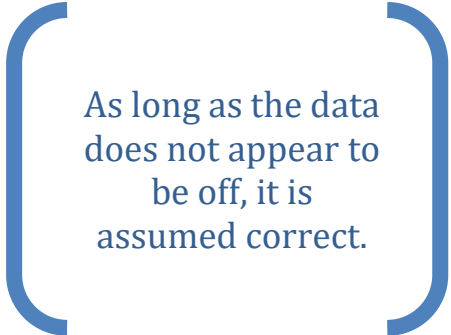
The Unknown Inaccuracy Data

The costs of poor-quality data are far greater than most organizations realize because inaccurate data often goes unrecognized. The hidden costs of unknown inaccurate data include the negative impact from using these data to make decisions that drive the organization. If the results of a data query are incorrect, how does one know the results are incorrect, especially if the answer was in an acceptable range? It is only when the distortion becomes obvious that it gets noticed. Organizations are not only challenged by their incorrect data, but even more with their ability to recognize it.

While most organizations recognize the need for data accuracy and quality, they attempt to "clean" or fix it without truly understanding what is "right." This is especially true when the data lacks business meaning, definition, or context. Unfortunately, our experience indicates there is an unconscious consensus in organizations that whatever "looks" right, must be right. For example, if a numeric query result was 10 times more than expected or character based then the data would obviously be questioned. But if the results are anywhere in an expected range where it "looks" right then the data's accuracy and quality is not questioned. It seems to be human nature to believe that whatever comes out of a computer is correct, similar to belief that whatever gets reported in the news is correct. The flawed assumption that whatever "looks" right must be right is one of the biggest data fallacies today that underlies our data challenges.

Data quality statistics account for only the data that is obviously, or proven, incorrect. As long as the data does not appear to be off, it is assumed correct. When questions are asked of the data, most assume the answers are correct, as long as the data does not look obviously incorrect. Many blindly trust their data merely on an assumption of it being correct, not really knowing the truth. Therefore, within what is assumed as good quality data there is a percentage likelihood of error.

> As long as the data does not appear to be off, it is assumed correct.

Unknown inaccurate data is a significant data challenge to most organizations. Data accuracy is a fundamental requirement for good decision-making and the unknown, inaccurate, and low-quality data used to make decisions is a great liability. This includes the loss of business confidence, lost opportunity, and compliance and security risk. These problems are multiplied as the unknown poor quality data is used and reused to plan, operate, predict, and steer business organizations.

Redundant Data

Redundant data results when data is intentionally or unintentionally captured and stored in different places. This includes both duplicated data and data representing the same business thing created independently for different purposes and applications. Inevitably, duplicated redundant data quickly becomes inconsistent data where the different copies of the data result in different values that are thought to be the same.

We find redundant disparate data is the norm in most organizations today and it is one of the biggest causes of data quality issues because it inevitably becomes redundant inconsistent data. Redundant inconsistent data is very costly to an organization, creating technical debt—each redundant data element requires additional code to input, move, and manipulate per system. There are also maintenance costs to support the redundant element in each system. Multiply this by the number of redundant data elements and the

hard cost is significant. When we include the additional human capital (person-hours wasted searching for the appropriate data source and linage from multiple sources of the same data), systems, storage, and network costs, as well as the intangible cost from the poor quality data inherent to redundant disparate data, the technical debt is huge. But an even greater cost results when there are many different versions of the same data, but only one version can be used to make a critical business decision. The effort and cost of determining which version is the correct version is staggering.

Another issue of redundant data is that the discrepancies are never resolved the same way: each department has a different answer for the same question. Inconsistent data issues can be elusive due to the departmental focus typical of most organizations; organizations where many departments are satisfied with their data silos and their versions of the data. This silo mentality has driven the proliferation of departmental data marts to be their business intelligence (BI) solution further supporting the silo redundant disparity data mentality.

Inconsistent data is one of the biggest culprits of failed analytics. The nature of analytics is like peeling an onion—as soon as you remove a layer or answer a question, there is another; one answer leads to another question and so forth. Analytical questions typically require data details that cross business areas (and their data silos). The deeper the question, the more holistic the data must be to answer the question. This requires consistent data from across the organization. It is nearly impossible for most organizations to address redundant inconsistent data that crosses departmental or functional levels. To do this typically requires an organizational shift to resolve the inconsistent data that exist across the organization. Our experience indicates that when there is an overall lack of a holistic data approach across the organization, it is far easier for organizations to continue to propagate inconsistent erroneous data. Because there is no "enterprise" focus, there is no requirement to reuse data and the organization lacks understanding the need for, or importance of, a holistic view of the organization.

> Inconsistent data is one of the biggest culprits of failed analytics.

Data-Information Overload

Data-information overload is a costly data issue facing most organizations. Organizations are continually building data structures and data systems to compensate for poorly defined or understood data, non-scalable and non-flexible systems, and poor quality data. This scenario leads to data-information overload where the volume of unusable data structures severely affects the utility of the usable data as well as the performance of IT systems. Data-information overload also puts an organization at risk in regard to compliance and information protection, where the unused data structures may contain sensitive data or the sheer volume of data hinders information discovery and protection.

The cost of data-information overload is significant as it includes the wasted time, labor, and productivity to support the unusable data structures, as well as additional system resource cost: the additional hardware and software to support the unneeded data and the additional human resources required to maintain all of the extra data structures. The greater costs are the intangible costs of customer dissatisfaction, lost opportunity, and erroneous decisions made due to data-information overload. Many organizations operate under the "keep everything" mentality as they continue to buy more storage capacity and technology. The "keep everything" philosophy fuels the data-information overload. Addressing data-information overload starts with identifying and properly addressing the lifespan of the data.

Inaccurate, Inadequate, or Missing Meaning

Most organizations have a significant amount of data that is undefined or inaccurately or inadequately defined. One of the main reasons for this is a lack of business connection to the data. When organizations skip business data requirements and data models, the documentation of the data, including business meaning of the data is compromised. Missing or inadequate data names and definitions result in the misuse of data leading to many additional data issues.

When data definitions are missing, technologists often create "dart board" or best guess data definitions. Though this sounds a bit dramatic, the dart board guess method for data definitions is becoming the norm. Dart board definitions are often more dangerous than no definition at all, because definitions add trustworthiness to the data. Data names and definitions formed by this "dart board" method make the data appear to be purposefully architected and designed. This gives the data users a false sense of trust when in fact there is often a high probability of error.

> When data definitions are missing technologists often create "dart board" or best guess data definitions.

Even with properly named and defined data, a widely used practice is overloading a data field. Overloading takes place when a data field is used for additional types of data than what the field name or definition signifies. This is typically done to circumvent the need for expensive system enhancements. Multiple users and/or departments can overload the same data field, each having their own cryptic translation of the multiple types of data placed in the overloaded data field. Time is also a key factor as the various types of data in the overloaded field can become outdated with no one remembering its purpose. Inevitably, the overloaded data fields collide with another department use of that field resulting in multiple data issues.

Because of the data overloading practice, the name and/or definitions of data fields are not always what they appear to be. This is a widely misunderstood fact about data that
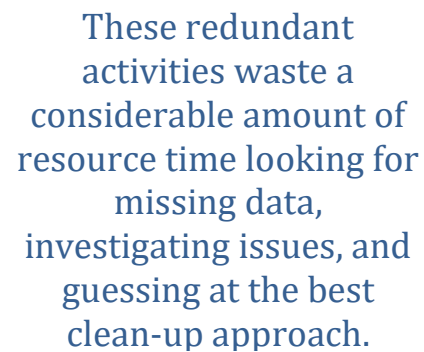
becomes a huge challenge for the automated data tools that depend on the names of data fields to effectively operate. Overloaded data fields can also be a significant challenge for information protection and security because they may contain sensitive information that the field name or definition does not indicate. For example, when a credit card number or social security number is placed in a customer name or some other "non-PII" field.

Management of Data

In most organizations there is a false sense that the data is being managed because the technology systems that house and process the data are being managed. Data technology management is not data management. Unfortunately, very few organizations today practice formal or intentional data management. We see organizations typically operating "data management" in a reactive mode with a tactical approach. With this approach data issues are addressed as they occur and only when they may negatively impact business operations. This boils down to "data management" on the fly in production as an emergency "Band- Aid" to fix a symptom but not address the core data problems. The underlying data problem continues to create data issues requiring "cleaning" week after week.

These reactive data management activities are costly and take place in many different departments, often with each department addressing the same data issue. These redundant activities waste a considerable amount of resource time looking for missing data, investigating issues, and guessing at the best clean-up approach. These reactive data management activities are typically outside the normal job activities of a department, and thus rarely accounted for. Without a central-coordinated and planned proactive effort, many organizations never realize the total cost to continually chasing their data problems.

Organizations rarely address the root cause of their data issues. Continuous project urgency and tight budgets drive a quick-fix (where the symptoms are quickly band-aided) rather than performing extensive root-cause analysis to address the issues. These "Band-Aids" often end up causing many more data issues than they "solve," leaving little time, money, and resources to strategically manage data and proactively get ahead of the issues. Real data management needs to be proactive and preventative.

> These redundant activities waste a considerable amount of resource time looking for missing data, investigating issues, and guessing at the best clean-up approach.

Even organizations with "formal" data management policies, standards, and processes often end up "practicing" data management in theory only. Proactive data management activities rarely take place in most organizations. The urgency of timelines and tight budgets pressure IT departments to bypass data management processes, sometimes without even realizing it. Despite formal documented data management processes, standards, policies, strategies, and governance being in place, nearly every project is still granted an exception justified with the intention of coming back and fixing it someday and doing it right the next time. But "someday" never arrives and there is no next time. These organizations fool themselves into believing they practice formal data management and then wonder why they continue to have data issues, questioning the effectiveness of data management as a practice rather than questioning the effectiveness of their data management execution. It takes time and commitment to get and keep the data right. The irony is that it takes more time and resources to get the data wrong.

> But "someday" never arrives and there is no next time. These organizations fool themselves...

Regulatory Compliance Data Challenges

The new regulatory and corporate governance requirements are creating multiple data challenges for organizations. Keeping data secure and private, as well as addressing a variety of compliance mandates, requires an accurate understanding of the meaning, location, ownership, source, and use of the data. When organizations lack meaningful data definitions and have data spread across and repeated in many systems, addressing compliance and keeping information secure becomes much more complex, if not impossible. It is a complex undertaking to identify data to be secured and/or protected with the lack of a data inventory and meaningful data definitions and the existence of redundant data everywhere. Even when the data to be secured and/or protected is identified, without healthy data management in place, securing and protecting the data is still extremely challenging due to the lack of standards, people, and processes.

Appropriately addressing compliance and securing information requires accurate accounting of the appropriate data fields versus listing the name of a data fields on a form during an audit. A column name or definition within a database often does not guarantee the data in that field is a match. Also, data fields can often be overloaded. The lack of proper data management means full analysis of the data fields is necessary to properly address compliance and information protection.

Compromises to the security of an organization's data, exposing personnel information is very serious as it has the potential to destroy an organization. Many financial and health organizations have been forced by regulatory compliance to address their inaccurate data issues, but most organizations continue to struggle with addressing the underlying

problems causing these data issues.

Acceptance of Data Issues as Normal

One of the biggest challenges regarding data issues is the acceptance of poor and inconsistent data as the status quo. This is typically based on the belief that this is a normal cost to doing business: a little low-quality data is either inevitable or harmless. Organizations typically underestimate the percentage of their low quality data, do not account for its hard dollar cost impact, and ignore the intangible costs. The flip side of this point is many organizations have either decided or have fallen into a belief that achieving data quality is impossible with the cost far outweighing the benefits.

Organizations across industries tolerate bad-quality data because they are unaware of the true cost of their data issues, and thus do not see an economically desirable reason to change. This is especially true for profitable organizations. Most do not see an alternative. Data may be one of the only corporate strategic resources where practices like this are acceptable. This acceptance has led to lack of urgency and focus on the data assets, thus continuing to fuel the enterprise's data challenges.

Key issues

What Caused Our Data Challenges

Defining our data challenges is important, but in order to address them it is essential to gain an understanding of what caused them. Our data challenges (i.e. poor quality, redundant, undefined, overloaded, etc.) did not just magically appear. Many factors have caused and continue to fuel the data mess. Unfortunately, even if there is a recognition of productivity drain from the data chaos, the focus continues to be on the symptoms rather than the fundamental data problems. It's like playing the role of Hans Brinker and putting one's fingers in the holes of the data dyke. Addressing data symptoms may be fine as a temporary fix, but there are far more data holes than fingers. We must understand the root cause of our data problems so we can get our data right and mitigate this data chaos.

> Unfortunately, even if there is a recognition of productivity drain from the data chaos, the focus continues to be on the symptoms rather than the fundamental data problems.

There are eight primary causes of our data challenges:
- The Computerization of Data
- Technology Driven Data
- System Development Practices
- Silo Functionality – Lack of Enterprise Focus
- The Tools Technology Myth
- The Quick Fixes
- Rapid Data Growth
- Unmanaged Data Aging

Computerization of Data

Before the evolution of technology, organizations stored much of their data on paper as records. These records were managed by the people who created and used them. With computer automation and electronic data storage, it became much more efficient to capture, store, and process data electronically. Much of what had not been practical to capture became a matter of a few key strokes. Technology has also made it easy to move and or copy these data records. With all the wonderful advantages technology offered came many data challenges.

Computer automation separated the data from the business. The people who create and use the data no longer manage it as they did when it resided in file folders. In the early days of computerization, the separation was not significant because there was only one system

and everything—systems, applications, and data—was supported by a few people who interacted daily with the business users. Data issues existed, but they were manageable. With the evolution of technology came the proliferation of systems, each with its own data, multiplying the data issues exponentially. The business became completely separated from its data as data became part of computer technology managed by the technology team.

Technology Driven Data

In nearly every organization the responsibility of data is placed under the IT department as a natural result of the evolution of information technology within organizations. Computerization began with automation of basic business functions which were previously being performed using calculators. As the computers became more sophisticated, so did the automation of business functions using software applications. The focus of IT departments has always been on the technology automation of business functions and continues to be the focus today. Data has traditionally been viewed as part of an application. Only in recent years with the evolution of systems designed for usage of data (e.g., the data warehouse, decisions support, BI, Big Data) has there been an increased focus on data—but only from a single data usage viewpoint. Even data focused systems are technology driven, rather than data driven. When the data is driven by technology, data systems are developed by creating table structures based on their support of the technology (i.e. coding methodology, hardware platform).

> The focus of IT departments has always been on the technology automation of business functions and continues to be the focus today.

In contrast, a data-driven system starts with a focus on the business the data represents. The data is architected and modeled in support of the business information needs and the need for data integrity. The next step is creating a physical design to balance the information integrity with the limitations of the physical environment. Technology should be an enabler of business capabilities rather than drive the business.
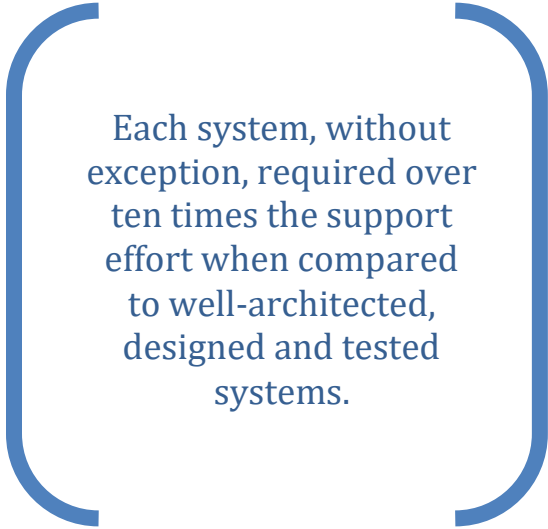
Technology-driven data supports the organization's need to quickly produce software. The effect of this is little desire or need for IT departments to spend the time architecting, designing, and defining data. In many organizations software developers create table structures as they develop code, with minimal or no business data requirements. Developer generated data structures typically have limited column names, making them difficult to decipher and adding very little value to the data documentation.

The developer's data structures can easily be reverse engineered into a data modeling tool and made to look like designed and architected data structures. Often best guess business names and definitions are added, giving a false sense of the trustworthiness to the data.

Many IT professionals see this practice as an acceptable method to "architect" the data. A modeling tool does not magically transform non-designed data structures into data architected designs. Unfortunately, this method creates many data issues which add to the deterioration of data. This backward approach is now the norm in most organizations.

There is a significant long-term risk with skipping Business Definition and Data Architecture, Modeling, or Design and directly generating physical data structures from application code. To save time, developers typically create data structures that align with and support their application code methodology and platform performance considerations. This saves time, especially with an agile development methodology. Although the data structures support the code (and a shortened implementation timeline), the data integrity and technical data debt, as well as total cost of ownership far outweigh the time savings. The long-term total cost of ownership of a system is rarely a consideration, especially in an agile development environment.

On many occasions I have personally experienced the scenario where important business systems were created on a "temporary" basis in order to "discover" requirements, skipping all architecture, design, testing, and migration. Development was on a sporadic as-needed basis. Data was moved from multiple legacy systems into more powerful technology. In reality, none of these systems were temporary and they quickly became the production environment that my teams supported. Each system, without exception, required over ten times the support effort when compared to well-architected, designed, and tested systems. This was due to the continual additional requirements and emergency data fixes requiring additional human resources and costly hardware and software, none of which were ever considered or planned.

> Each system, without exception, required over ten times the support effort when compared to well-architected, designed and tested systems.

Systems that are technology focused and lack architecture, modeling, and design are more costly to maintain because they require continual adjustments, band aids, and enhancements from missing requirements, along with continuous effort to "fix" the data victims. With little or no data definitions, incorrect and missing data fields, missing integrity constraints, hard-coded inflexible structures, "fixing"

the data issues is much more challenging, if not impossible, requiring additional time money and resources. None of these extra maintenance costs and intangible risks are typically calculated into the cost of the project. Total cost of ownership is rarely considered in today's development methodologies. The intangible risks include the non-recognized data integrity errors which result when unknown erroneous data is used to make important business decisions and create flawed reports. It is very difficult to account for these intangible risks that result in a significant cost for an organization.

> None of these extra maintenance costs and intangible risks are typically calculated into the cost of the project.

A technology focus drives the creation of database structures that support performance of specific code and functions. Future business needs and functionality are rarely considered. This focus on performance leads to rigid data structures that cannot be changed without considerable costs. Because of this, when changes are required (and they almost always are), organizations create redundant data structures rather than redesigning existing structures. These redundant data structures soon get out of synch, resulting in data quality issues and increased data overload. This in turn decreases productivity and increases costs and expenses due to the time spent questioning data, hunting for errors, correcting bad data, and redoing work. Even worse, if the bad data ends up on reports and analytics or is given to customers, then the cost rises dramatically. Redundant data also adds costs in terms of system functionality and maintenance. With the increased data overload, the greater the volume of data, the more difficult it is to find the data needed, and the more hardware is needed to house and support this additional data.

System Development Practices

System development processes practiced within organizations have contributed to our data challenges. Most data systems' development projects do not include consideration for how the data will accurately represent the business. Also, organizations shorten or eliminate data usage definition or business requirements steps in the typical agile development methodologies. In addition, many compromises are made to physical data structures during the design step due to technology limitations and performance requirements, further compromising the data. The data is yet further compromised by all the "temporary" fixes applied during emergency situations—situations driven by the minimized business requirements and design steps. The end-result is more unknown distorted data.

When business requirements are used to design new data structures, they are often given in the form of a "laundry list" of data attributes. By its very nature, a list of data items is out of context. Without context, the relationships that represent the complexities of the real-

world business are missing. Everything in a business organization is interrelated and dependent; nothing exists on its own. Taking data out of context severely compromises its meaning and integrity. The importance of this concept is often not understood.

As technology rapidly advances, organizations continually upgrade systems. IT departments, strapped with tight budgets and timelines, tend to simply "lift and move" the data from an older to a newer technology. Organizations often shorten or skip business requirements while replacing a system because they believe it will add to timelines, especially with the conversion processing from the old system to the new. If there are any business requirements they are generally used for documentation rather than the design of new data structures. Using the data structures from an existing data source appears to be a much quicker method. Most IT departments assume that the newer technology will fix many of the data issues and limitations of the old system. But in reality this method only moves the data disparities—known or unknown—forward.

Many organizations are now using packaged solutions to automate common business functions. Each of the packaged software solutions has its own proprietary generic view of data needed within the application. The organization's business functions are distorted to fit into the package's generic view, adding yet another version of the truth. These packaged solutions typically use the data from the systems being replaced in a "lift and move" fashion. Despite efforts to "clean" data during conversion, many of the old systems' data issues are moved forward because of the unknown nature of these issues.

The data distortions grow exponentially each time a new system is implemented using the previous system's data. It's similar to taking a photo of a photo of a photo and on and on until everything becomes gray. Each time the photo is copied we lose resolution. Please see the diagram below:
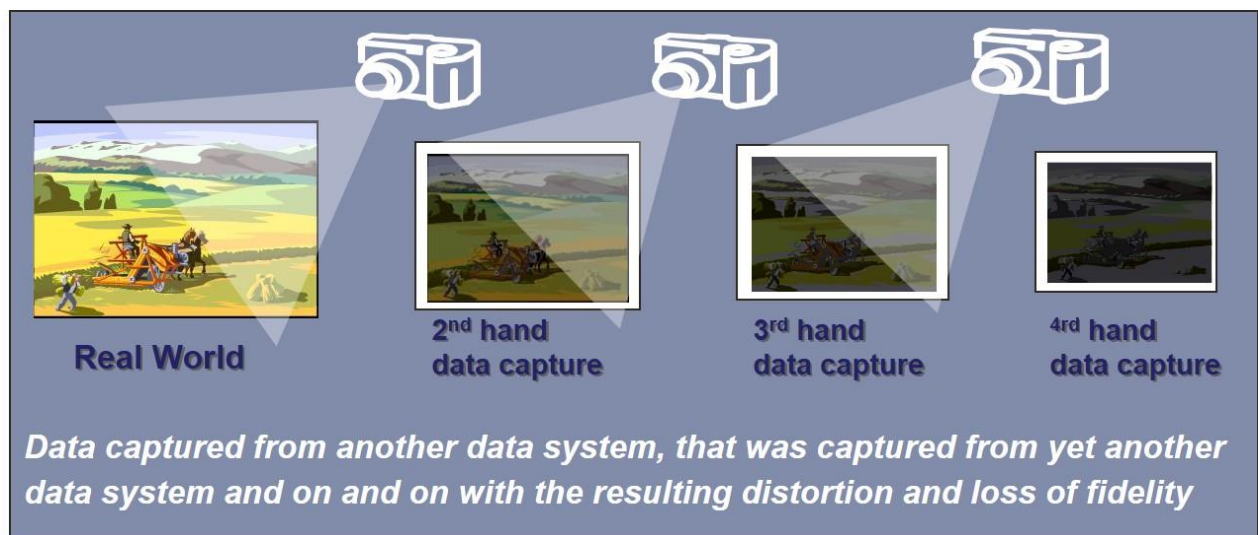
Figure 2: Example of Loss of Data Fidelity

In terms of data, the "resolution" is the connection between the business and the data. The result is a growing disconnect between the data and the business reality it represents. As data moves further away from the business reality, the organization suffers from a growing number of data problems.

Silo Functionality
Organizations may start off with one version of the truth but quickly end up with multiple versions. Most business systems were never designed for integration. Each is built one project at a time with defined boundaries that support a specific business function. Most projects use different people, funding, technologies, assumptions, business rules, and, oftentimes, different data sources. Rarely do these individual projects consider other systems and projects. This all eventually leads to silos of functionality and data, as well as multiple version of the truth.

Many inconsistent data challenges result from different departmental processes with overlapping functionality. The overlapping functionality requires the same data, but often each applications' data is obtained from different sources or "customized" for the particular business function. This can quickly result in inconsistent data if it is not managed properly.

The major driving force to this approach is focus, control, and reward at a departmental level. A departmental focus is reinforced by department level goals, incentives, and performance objectives. These foster the idea that each business unit is more or less independent. Many managers easily forget who the real competition is: each department works as if it were an independent business that competes for the same scarce resources. This reinforces departmental silos of data and discourages teamwork, which is fundamental for data integration and sharing.

The Quick Fixes
Technology failures threaten business operations more frequently as the data becomes more and more distorted. IT organizations typically resolve these situations with a "quick fix" that the organization views as essential due to the potential loss of business. Many of these "quick fixes" are a compromise to the data structure and thus the data integrity. The known compromises always come with the famous last words of "this is only temporary, we will come back and do it right someday." Unfortunately, someday never comes and the data remains distorted. Adding to this challenge is the perceived lack of time to document how these fixes distorted the data. This practice leads to an even greater distortion of the data from the business reality. There is always a price to pay when urgency wins over quality.

> The known compromises always come with the famous last words of; "this is only temporary, we will come back and do it right someday."

Rapid Data Growth

TechVision is seeing data stored within organizations doubling every 12-18 months. Data's growth is due to both the organization's continual demand for information and the continually richer content available from documents, media, graphics, and even sound. Technology is enabling this growth as the cost of technology continues to drop while its capabilities increase. Also, the evolution of analytic capabilities has made many organizations aware of the need for greater and greater understanding from the questions asked, all leading to requirements for more data detail. There are also all the new technology devices that capture and use more data anytime, anywhere.

Rapid data growth fuels the data overload challenges facing most organizations (please see Data-Information Overload section). At the same time budgets and resources are tightening in the face of the world economic competition. A significant data challenge from this is the inability of the current data management and business techniques to keep up with this growth. Although data technology is managed, the actual data is not managed in most organizations, adding to the data challenges resulting from rapid data growth.

Unmanaged Data Aging

An organization's data is organic in that it has a lifespan: data continually ages. Data continuously changes as it ages in both state and value. Because of this it is essential that an organization understands the data's lifespan. The theory of entropy proposes that everything moves toward a state of greater disorder and it takes intentional, continual energy to keep it in order. Data too continuously changes and heads toward disorder resulting in significant data challenges, especially when it is unmanaged.

One of the biggest challenges of data aging is realizing when the data has reached a point where it is no longer of value and therefore should not be kept. The volume of data and little or no data management practice makes it nearly impossible for organizations to know when data's value is such that it should be deleted. In a vicious cycle, this continues to add to the data overload and its data challenges.

The Tools Technology Myth

In spite of the growing number of tools and technology touted as the "magic" solution to data challenges, the problems continue to grow as rapidly as the data. As mentioned earlier in this report, organizations continue to do the same thing (buy into the next "Silver Bullet" tool, technology, or magic data fixes), continually hoping for different results. This continuous cycle of believing the next silver bullet will fix the data issues distracts organizations from considering or addressing the real data problems. Tools and technology are never "the solution"; they can be used to automate solutions. The data challenges facing organizations today are the results of the broken fundamentals of data and require a fundamental change in our thinking and approach.

Discussion

The Fundamentals of Data

The fundamentals of data are the underlying principles, or beliefs, that are taken for granted—these form the basis for our approach to data. Fundamentals are the central or primary rules and principles on which something is based, and thus are assumed to be correct. The fundamental beliefs of data form the foundation for how we as an industry view and approach data. All of our data practices are based on these fundamentals.

The fundamentals of data are broken. When the fundamentals are broken our approach to data is flawed, resulting in flawed data systems and poor data-quality which severely handicaps our full data usage.

The five core broken fundamentals of data are:
- The Fundamental Understanding of Data
- The Connection of the Data to the Business It Represents
- Representation of the Real World using Data
- Real World and Its Data Translation into the Computer World
- Data as a Business Asset

The Fundamental Understanding of Data

Most consider data as part of computer technology: the electronic symbols or tokens used within computers, networks, and storage devices, to support software applications. In nearly every organization data is placed under the charge of the IT Department. Not only do IT professions perceive data as a technology component, as do most business professionals. *Data is not a technology component.* Our fundamental understanding of data is flawed.

Data is a representation of the real world and has existed long before computers were ever imagined. The real world perspective of data has been buried in the avalanche of technology throughout the history of computerization, where the focus has always been on the automation of functionality. Data has rarely been viewed outside of applications. Understanding what data really is—a representation of the real world—outside the confines of computer technology is fundamental to "getting the data right."

> Understanding what data really is, a representation of the real world, outside the confines of computer technology, is fundamental to "getting the data right."

Data is a representation of the real world: the tangible or intangible things, events, and relationships. As a representation, data stands in for whatever it represents; it serves to

capture the essence and account for the known or inferred properties of a real world thing. Data captures the characteristics of the real world that are considered important, which includes all of the interrelationships of things and events.

Everything in the real world is interconnected and interrelated, thus relevant to its surroundings and its relationship to its surroundings. This concept is supported by the basic theories of physics. Nothing exists on its own in the real world. Everything derives meaning from its relationships. In other words, context gives meaning to everything. Removing or changing the context changes its meaning. Therefore, capturing the context or relationships of the real world things and events is essential for success when it comes to understanding data.

Data as a representation of the real world facilitates human understanding and knowledge by recording the past reality in order to analyze the past as well as predict the future. Knowledge of the real world can be gained from the data that represents the real world. The validity and usefulness of the data is dependent on the accurate alignment of the data to the real world that the data represents.

Because data has been considered a technology component, tools and technologies have naturally been the focus to solving our data challenges. After all, technologists naturally apply technology solutions to problems. Furthermore, when data is seen as a technology component, it is not treated as its own entity separate from the applications. This creates silos of data within the application driving more data issues.

> This technology view of data has greatly hindered the concept of data as a business asset.

This technology view of data has greatly hindered the concept of data as a business asset. Data is an important business asset and not a technology component. As an asset, data is independent of technology and must be viewed and treated as such. The fundamental misunderstanding of data as a technology component is at the core of the broken fundamentals of data.

The Connection of the Data to the Business

With the fundamental misbelief that data is a technology component, the connection of the data to the business is weak or missing in most organizations. The concept of data as a representation of the real world business organization requires a clear understanding of the connection of the data to its business meaning—the connection of the data to what it represents in the real world.
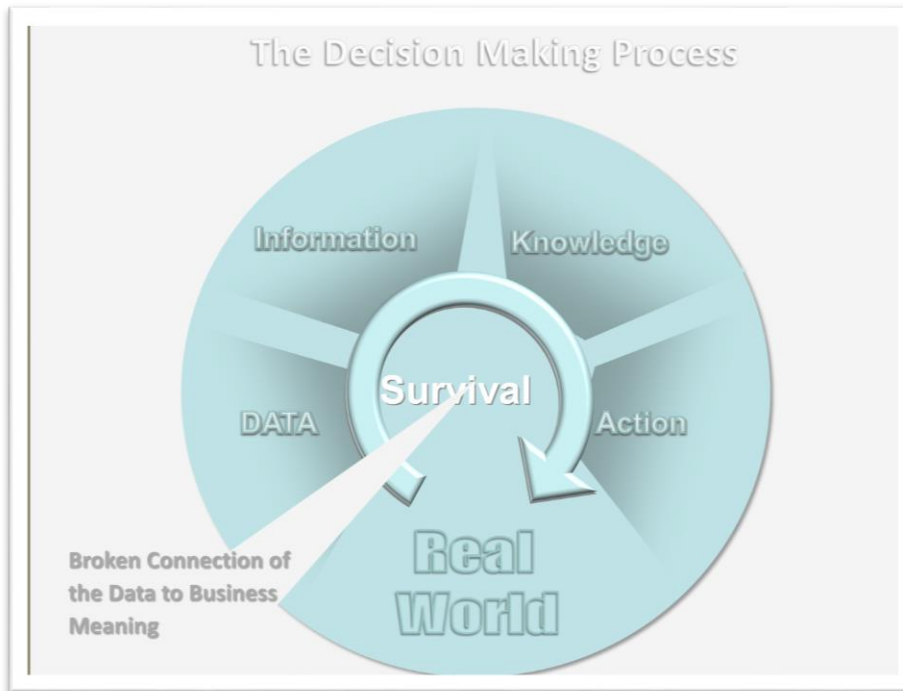
Figure 3: Broken connection between the Data and its Business Meaning

In most organizations data is closely associated with a software application: a usage view of the data. What data represents and its usage are not the same thing. Similar to the real world, where anything in a business organization can be used in multiple ways and within many business functions and applications, a single usage of data gives only partial view of what those data represent in the real world business. The data lacks the holistic view of the business. The holistic view of the business organization that the data needs to represent identifies the things and events important to an organization, as well as how these things are related.

Having a holistic view of the business will give a much needed business perspective of data—bridging the gap between business and its data in a consistent manner. This is the key to turning data into a real business asset. Similar to a technology infrastructure, a data infrastructure is necessary to effectively support the data assets in a consistent manner. A planned and architected technology infrastructure consistently supports the technology that stores and manages the data. The data infrastructure is architected and built based on a holistic view of the business organization that the data represents. To complete the data infrastructure, each usage of the data is linked to the holistic view of the business organization, tying all of the organization's data to the business. The gap between the business and its data is a big issue in most organizations today, making this a core reason the fundamentals of data are broken.

Representing the Real World using Data

Because organizations have historically viewed and treated data as a technology component rather than as a representation of the business organization, a consistent methodology for representation of the real world using data has never been established or formalized. Actually, organizations have not even recognized the need for a consistent methodology. This has fueled many of our data challenges—especially those resulting from inconsistency, redundancy, and integration.

Most organizations capture or create data on an as-needed basis as each application is built. Not only does each application determine what data it needs, but each person in an organization has their own perception of what data needs to be created and captured. The situation is made worse by the fact that each application is only a partial view of the data, because each application is only a single usage view of that data. The end result is: the same real world business thing or event is represented by different data, as well as each with a different method of representation, most of which are completely random. The majority of data within systems today is lacking any rigor or consistency regarding how the data represents the real world business things and events. This results in inconsistent, redundant, and disparate data.

As technology evolves, organizations continually replace systems containing legacy data, but often "lift and move" the legacy data into the new technology. This simply moves any inconsistent representation forward! Even when new systems are created using various forms of "data models," the data model does not define a consistent method for how data must represent the various types of business things and events needed by the application. Even if a data model were to ever include a method of representation, this would only focus on a single usage of the data. A method of representation of the real world business using data must be based on a holistic view of an organization rather than on a single usage view.

A formal methodology regarding how data will represent the things and events of a business organization must be based on a holistic view of the business organization, including a fundamental understanding of data as the representation of the real world. Each important type of thing in a business organization (e.g., time, people, items, events), must have a standard for its representation, including a standard set of real world dimensions. This is essential for consistency. Some of the questions that need to be asked by the data professional when designing a methodology include:

> A method of representation of the real world business using data must be based on a holistic view of an organization rather than on a single usage view.

- How do we effectively use data to represent the real world?

- Can data adequately capture the real world? Are their limitations?
- What are the challenges to capturing the right data?
- What facts are necessary for each category or type of business thing to represent it appropriately by the data
- What constitutes an accurate representation of that real world thing?
- At a specific level, what facts do we represent and what method to we use to determine the appropriate facts to represent?
- How do we determine the effectiveness of the data representation?

The data will be far more consistent, complete, and accurate—thus much simpler to integrate, define, understand, and utilize within applications—when the organization uses a formal data representation methodology. Standards, structure, and consistency are critical to achieving quality data. No one application or usage of data should control what data is necessary to represent business things and events. This needs to be determined prior to any usage of the data. The data representation of real world things and events is not the same thing as the functional usage of that thing or event: any business thing can be used within many business functions. The methodology needs to be holistic and applied across the organization.

> No one application or usage of data should control what data is necessary to represent business things and events.

Real World Translation into the Computer World

Along with the other broken fundamentals of data, comes the apparent lack of recognition by organizations of the necessary translations between the real world and the data. Understanding the difference between the real world and its data representation in relation to the limitations of computer storage and processing is a fundamental concept that is core to the proper architecture and management of data assets. We must clearly understand the differences and proactively address them.

The translation flow begins with the real world to be represented by data. The next step is preparing and transforming the data representation into the electronic or digital world down to the machine level (i.e. 0s and 1s). At a later point the zeroes and ones are reassembled and transformed back into the data, providing information and knowledge about the real world. This process begins and ends with the real world, thus it is critical that throughout the translation fidelity to the real world remains intact. Loss of fidelity is one of our biggest data challenges.

Intertwined with this translation flow is human perception: no two people's perception of reality is identical. Each translation step (between the real world and computer and back) poses a risk to the loss data fidelity. To minimize risk, the translation process and its potential impacts must be clearly understood and proactively managed.

TechVision RESEARCH
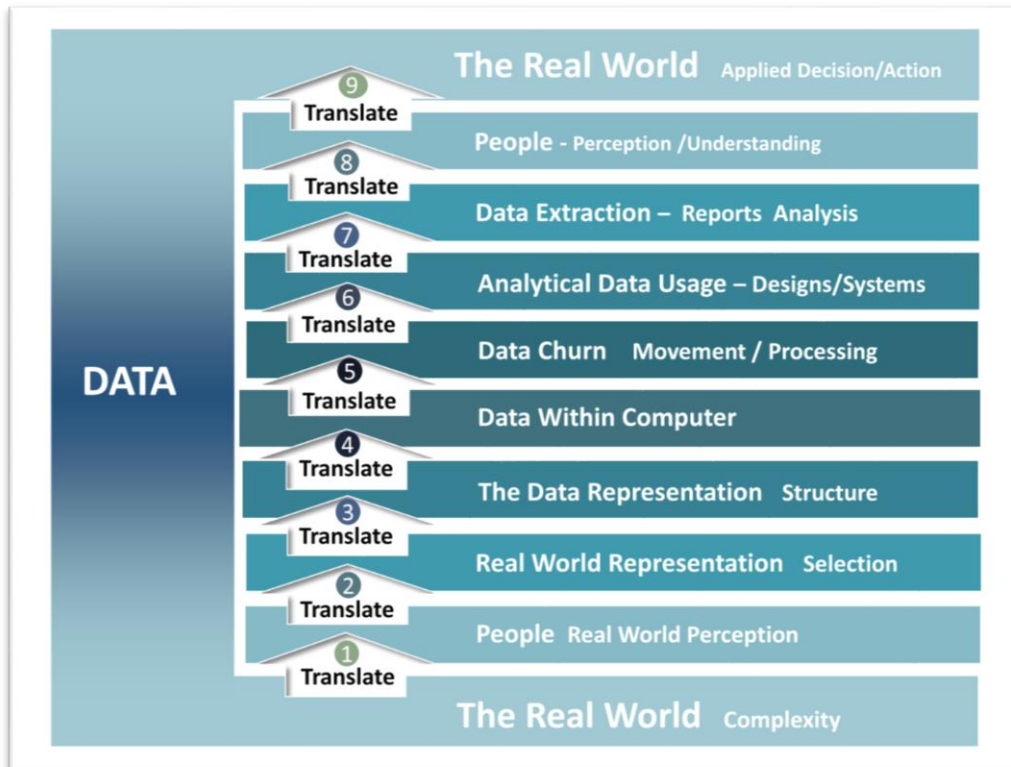
*Figure 4: Real World Translation into comput*er and back to the Real World

The nature of the real world is very different from the nature of computers and computer storage. The real world is infinitely complex where computer systems are finite. Humans and computers use data and information differently. Humans process information through their senses, by induction (deriving general principles), and by deduction (the process of reasoning). A computer is an advanced machine created by humans as a tool to perform "thinking" work in a predetermined logical manner. When the differences are not recognized, well understood, planned for, and managed, the result is loss of data fidelity.

Most organizations morph the real world (or its data representation) to fit the computer rather than carefully translating the data to retain its real world context and fidelity. This creates many of our data challenges. In other words, the data is distorted to meet the limitations imposed by processing and storage technology. Thus data becomes technology-driven. With technology-driven data there is *always* a loss in context and fidelity.

Maintaining the business context and the fidelity of data requires correctly understanding the real world business: consistently capturing the real world with data, and then translating the data so it can be organized, stored, and processed by technology in a manner that supports the optimal use of that technology. This translation must be purposefully architected, designed, and documented so that the data from all of the various systems will be translated back to the real world in a holistic and consistent manner that

maintains fidelity. Critical to maintaining fidelity is including all of the data's relationships and interdependencies. Rather than depending upon the technology system to maintain the data's fidelity and translation—which has never worked—the translation must be managed independently from the technology.

It is critical that the real world and the computer world remain independent, and the interface and translation between the two must be carefully understood, designed, and managed. In order to retain its real world validity (i.e. context and fidelity) as data is translated into and out of various technologies, data must be considered independent of the technology. In the same manner the technology needs to be optimized rather than restrained and limited by the data. The independence that both the data and technology require can be achieved and maintained using a carefully designed interface (i.e. translation), which is independent of both the technology and the real world data.

Data as a Business Asset

Data understood within the context of technology rather than data as a representation of the real world business, naturally, makes it difficult or impossible for most organizations to accept the concept of data as a business asset. Although the concept of data as a business asset is given lip service and may even be included in the strategy, data being treated as a business asset is not a reality for most organizations. Treating data as a business asset requires first understanding data as a business asset. This requires understanding data as a representation of the real world business organization rather than as a technology component of the business organization. When organizations see data as a representation of the business, the data almost becomes the business, making it critical to the business' success.

A widely accepted definition of an asset is: something of value that is measurable and helps an organization achieve its strategic objectives. Data clearly has an economic value to an organization because if it's destroyed, lost, or stolen there would be a significant cost. However, data can only be considered an asset when it is formally managed. Unmanaged data is not considered an asset. Financial Accounting Standards Board (FASB) has another definition of an asset. The FASB considers any economic resource of an organization to be an asset. According to the FASB definition, data should be listed on a balance sheet with the other assets, but data is rarely, if ever, listed on a balance sheet.

> Although the concept of data as a business asset is spoken about and may even be included in the strategy, data being treated as a business asset is not a reality for most organizations.

Data is the raw material of information—similar to the raw materials used in manufactured goods. Both the raw materials and manufactured goods are considered assets. Data and

information can be viewed as an asset similarly to manufacturing assets. Data as raw material goes through value-added processes and into finished goods or usable information. As with goods manufacturing, this transformation process requires capital investment and human labor. Likewise, data-information requires management, much like any other asset. Similarly, management should include quality-control inspections before usage or distribution of any data assets. Finally, the same just-in-time principles applied to manufacturing inventory can similarly be applied to the data-information assets.

To manage anything effectively, it needs to measured and assigned a value. Typical assets are assigned a dollar value, along with how that dollar value is measured. Data clearly possesses value though assigning monetary value to data is a concept that few organizations are embracing. Organizations can perceive the value of data as an actual value (i.e. based on a measurement) or it can be defined by the risk-based cost if that data were to be lost, damaged, or stolen. The more data is used the more the risk of loss increases, as does the data's value.

Management of any type of asset requires a carefully planned infrastructure involving processes, rules, people, tools, oversight, and, most importantly, funding. The appropriate processes, dedicated people, and funding are important for the success of data asset management. The appropriate asset management processes need to be in place because the management of any asset is primarily a set of processes. But without appropriate funding and people dedicated to Data Asset Management (DAM) processes, it will quickly become just another good idea that bites the dust.

In order to properly manage anything, it has to be first identified, defined correctly, inventoried, and then valued. Not all assets carry the same value and thus require the same degree of rigor. In most organizations today the identification of data is extremely challenging due to the lack of business definitions, data ownership, appropriate data modeling and design, as well as the high degree of redundant data.

> But without appropriate funding and people dedicated to Data Asset Management processes, it will quickly become just another good idea that bites the dust.

The majority of initiatives today involving managing data as a business asset get twisted into a tools or technology sales pitch. Many of these tools based methods overlook the most essential step: getting the data right in the first place. If the data was never right in the first place, then all the asset management methodologies in the world cannot make it right. Since data is a representation of the business organization used to track, manage, and predict the business, an accurate representation is key to treating data as an important business asset. The entire organization is dependent on a correct representation—as every single function

within the organization, including business data asset management, hinges on this.

Getting the data right is fundamental, but treating and managing data as a business asset is the critical key to keeping the data right. Data as a business asset is a fundamental data principle and practice that is missing from most organizations today. Data asset management concept, strategy, principles, and practice need to be established before we can fix our data challenges.

Recommendations

Fixing the Fundamentals

The fundamentals of data are clearly broken. Our basic understanding and approach to data is flawed. We need to fix the fundamentals; get the basics right first in order to address our data problems, and then build or buy viable data systems that can enable the business to gain full usage of its data assets. Until we fix the fundaments, no tool or technology will solve our data problems and the cycle of data chaos will continue to grow.

As each "new" data technology emerges, hits its peak of inflated expectations, and then sinks into disillusionment, more continue to emerge. Now that Big Data has passed through its hype cycle and has become a reality in many organizations, the emerging trends now focus on the usage of Big Data. Let the data challenges begin. Many organizations have come to realize that with big data they face the same challenges of small data, but on much a bigger scale—data quality issues, integration challenges, and insufficient data governance and management. All of the great data technology advancements will continue to fall short of expectations mostly due to the reality of our data—redundant, disparate, inaccurate, missing, misplaced, undefined, miss-defined, dirty, old, reused, misused, etc.

Technology advances have far surpassed the limitations of our data. Nearly all of IT funding, resources, and focused effort have gone into technology, with little or no attention to the actual data itself. After all the name is "Information Technology," indicating equal focus on the information and the technology. Unfortunately, most of the attention is on the technology. The technology is seen and touched, whereas the information is the fuel that drives the technology. The technology on its own has little value to the business. The disproportionate attention is similar to building a high tech, state of the art, precision automobile and expecting top performance despite using unrefined crude oil for fuel.

> Nearly all of IT funding, resources, and focused effort have gone into technology, with little or no attention to the actual data itself.

Data is an important business asset and technology is its enabler. We need to spend as much or even more time, effort, and resources on our data as we do on our technology or we will never gain the benefits promised by the technology. We must start by fixing the fundamentals of data.

There are five basic steps an organization can take to address the broken fundamentals (each will be covered in-depth in subsequent papers):
- Establish the Business to Data Connection using a Business Blueprint
- Create a Data Oversight Framework
- Establish an Enterprise Data Construction Practice

- Build the Data Asset Management Infrastructure
- Establish a Data Asset Management Practice focused on Enterprise Foundational Data

1. Establish the Business to Data Connection using a Business Blueprint

Grasping the necessity and importance of the Business-to-Data connection requires the fundamental understanding and embracing of data as a representation of the real world. The Business-to-Data connection is core to all of the broken fundamentals of Data; therefore, we first need to recognize the importance of the relationship between the data and the real world and understand why this is key to fixing our data issues.

Data represent the real world business organization—its things, events, and relationships (tangible or intangible). Data, along with the information gleaned from the data, can empower the organization, but only when the data accurately represents the real world things and events it was intended to represent. If the data does not accurately represent the real world thing or event, then every use of those data will be compromised. Thus, understanding and documenting the connection between the business organization and the data that represents it is critical for success.

The Business Blueprint is a proven method to understanding and documenting a Business-to-Data connection that serves as a foundation for connecting the data to the real world business organization: the organization the data is intended to represent. The Blueprint is a holistic informational diagram of the real world business organization defining the important things and events that the data must represent. To understand and plan what data is needed to optimally and effectively represent the business, the holistic blueprint of the business must be created independent of any data models or designs. The blueprint is not a model of the data, rather it is a model of the business the data will represent.

> The blueprint is not a model of the data, rather it is a model of the business the data will represent.

A Business Blueprint establishes the framework for a "data" infrastructure that can be used across an organization to support a holistic view of the data. This is essential for accurate data integration, analytics, reporting, security, and protection of the data assets. The Business Blueprint data infrastructure framework serves as a Launchpad for any new data system, purchased or built in-house, as well as a map for connecting all the various application data stores needed for any Business Intelligence (BI) functionality. The Business Blueprint's role in accurately connecting the data to the business it represents is core to fixing the broken fundamentals and getting the data right.

2. Create the Data Oversight Framework

The second step to fixing the fundamentals is building a Data Oversight Framework (DOF) to establish a "playbook" (the strategy, principles, policies, and rules for the information-data assets) along with the functions to implement and support it. The framework covers all of the data strategy components necessary to establish and orchestrate the data assets' well-being. The DOF is foundational for all other data strategy components (e.g. data governance and ownership, data security, data asset management, etc.). Developing the framework gets everyone on the same page and in agreement as to the direction, value, importance, and priority of the information-data assets.

Many organizations have attempted to launch an information data governance function and have failed. The major reason for this failure is the broken fundamentals, where the business-to-data connection and a Data Oversight Framework (DOF) are missing. After all, it is very difficult if not impossible to govern anything without a correct understanding of what one is governing, the rules to govern by, and the infrastructure necessary to execute the governance. Just imagine attempting to govern or control traffic without traffic rules, driver's licenses, law enforcement, basic street signs, traffic lights, and lane markers.

In every organization where TechVision Research has addressed data governance, the Business-To-Data Connection and the Data Oversight Framework (DOF) were missing. In every case, our first step to establishing a Data Governance Practice is standing up the Framework along with a Business Blueprint. With these in place the governance practice has functioned well and thrived.

> The major reason data governance fails is because of the broken fundamentals, where the business-to-data connection and a Data Oversight Framework are missing.

3. Establish an Enterprise Data Construction (EDC) Practice

The third step to fixing the fundamentals is establishing an Enterprise Data Construction (EDC) Practice for the development of data systems and structures to house the data. A holistic (i.e. city plan) approach is used to properly support data as a representation of the real-world business organization. The EDC Practice covers the identification, architecture, design, and deployment of data structures and systems across the organization including the organizations meta-information. The practice uses the Business Blueprint as the foundation for all of the organization's data and data structures. This forms a holistic data infrastructure that ties all of the organization's data systems together.

Making the Business Blueprint the foundation and applying the practice enterprise-wide helps to assure data consistency. The outcome of the Business Blueprint and EDC Practice is an extensible data infrastructure that may be used by data integration, Data Asset Management (DAM), information security, and Business Intelligence (BI) functions. This approach is essential for effective operation. Without a consistent documented data infrastructure, the ability to integrate, manage, secure, and gain intelligence from the data assets is difficult at best and impossible at worst.

> Without a consistent documented data infrastructure the ability to integrate, manage, secure, and gain intelligence from the data assets is difficult at best.

The EDC Practice includes multiple approaches to accommodate the various types of data systems and the range of data asset values. No two organizations are the same; therefore, the EDC Practice approach must align with the organization's maturity level, business objections, etc. However, the EDC Practice goal is the same for every organization: utilize the Business Blueprint to build out the data infrastructure in a holistic consistent manner in order to orchestrate the usage and management of the data assets across the organization.

4. Build a Data Asset Management (DAM) Framework

Getting the data right is only half the formula to fixing the fundamentals. The other half is keeping it right. Data is organic in that as it ages it continues to decompose and deteriorate. As with any other business asset, the data assets too must be managed, and in order to do this the organization must build a Data Asset Management Framework.

The forth step to fixing the fundamentals is creating a Data Asset Management (DAM) Framework by building the infrastructure necessary to manage the data as an asset. The Framework includes the methods, processes, procedures, and tools required to manage data as an asset and it utilizes the data infrastructure developed through the Business Blueprint and EDC Practice.

5. Establish a Data Asset Management (DAM) Practice focused on Enterprise Foundational Data

The fifth step to fixing the fundamentals is establishing a Data Asset Management (DAM) practice with an initial focus on Enterprise Foundational data. This practice applies proper asset management principles and methods to the business organization's most critical type of data assets: the enterprise foundational data asset.

Enterprise Data is typically the data needed and used by multiple business areas, departments, and/or functions within the organization and thus it carries significant importance (i.e. significant to an organization's operations: both tactical and strategic). Enterprise Foundational Data is highly valuable since it is both shared across the

organization and it serves as a fundamental building block for other data. It's important to note that not all foundational data is enterprise and not all enterprise data is foundational.

Correct categorization is key to success. This includes understanding the various types of data used within an organization and using the data taxonomy to categorize it correctly. This Taxonomy of Data has multiple uses, but for DAM it is used to apply a high level valuation metric to determine the appropriate level of asset (i.e. the data) management necessary for the various types of data.

Most organizations are experiencing massive volumes of data, making it impossible for organizations to feasibly manage all of its data as a business asset. Because Enterprise Foundational Data is the data with the highest value it must be the first to be managed as a business asset. Therefore, it is critical to identify the Enterprise Foundational Data.

There are clear benefits to prioritizing management of the most valuable assets first. Enterprise Foundational Data are generally not a large set of data but the management of it provides the greatest return on investment when compared to managing any other type of data. Enterprise Foundational Data directly affects nearly every business function. Once an organization successfully manages its Enterprise Foundational Data, then it can begin to manage its next highest valued data.

Conclusion

Data challenges have become one of the most underestimated challenges organizations face in the information age. The sheer volume, lack of quality, definition, and mismanagement of data are crippling organizations. These data challenges greatly hinder an organization's effectiveness, growth, and profitability, perhaps even leading to that organization's demise.

Without question, data is one of the most valuable assets a business organization owns, yet rarely do organizations account for their data as an asset of value. Most everything in an organization today is based on its data: from recording and operations to predicting, planning, reporting, and decision-making. Rarely are business decisions made on actual observations in today's world. Rather, these decisions are made based on the data that represents the business. Because of this, all critical business decisions rely on a belief in the data's credibility. Since everything in a business organization hinges on its data, accuracy is absolutely essential for an organizations optimal functioning and survival.

As discussed in this report, the fundamentals of data are broken. TechVision Research sees too many organizations in which the basic assumptions regarding data are either flawed or missing. Addressing the broken fundaments is the key to stopping the data chaos and addressing the underlying data problems causing enterprise data challenges. This requires organizations to first embrace the fundamental truth that data is not a technology component. Rather, it is a representation that serves to capture the real world things and events that make up the business organization. The validity and usefulness of data is directly dependent on the accurate alignment of data to the real world it represents. Accurate data yields reliable information and reliable information leads to accurate knowledge and decisions necessary for the organization's ultimate survival.

To overcome these challenges, organizations must fix the fundamentals, get the basics right before they can ever hope to get the data right and truly utilize its full potential. I had a basketball coach who would always say "if we cannot get the basics right (e.g. dribble, pass, and shoot) then we cannot possibly pull off the fancy plays it takes to win the games." This same principle holds true for data: if an organization does not fix the fundamentals, it will never be able to effectively clean, identify, integrate, manage, and utilize its data assets for even basic operations (e.g. reporting, analytics, security, and planning), let alone taking advantage of the full power of its data assets for true business intelligence, risk avoidance, predictive analytics, artificial intelligence, and data monetization. Information is truly powerful, but only if the data is right.

About TechVision

World-class research requires world-class consulting analysts and our team is just that. Gaining value from research also means having access to research. All TechVision Research licenses are enterprise licenses; this means everyone that needs access to content can have access to content. We know major technology initiatives involve many different skill sets across an organization and limiting content to a few can compromise the effectiveness of the team and the success of the initiative. Our research leverages our team's in-depth knowledge as well as their real world consulting experience. We combine great analyst skills with real world client experiences to provide a deep and balanced perspective.

TechVision Consulting builds off our research with specific projects to help organizations better understand, architect, select, build, and deploy infrastructure technologies. Our well-rounded experience and strong analytical skills help us separate the "hype" from the reality. This provides organizations with a deeper understanding of the full scope of vendor capabilities, product life cycles, and a basis for making more informed decisions. We also support vendors in areas such as product and strategy reviews and assessments, requirement analysis, target market assessment, technology trend analysis, go-to-market plan assessment, and gap analysis.

TechVision Updates will provide regular updates on the latest developments with respect to the issues addressed in this report.