CHECKLIST REPORT

2018

Five Data Management and Analytics Best Practices for Becoming Data-Driven

By Fern Halper

Sponsored by

sas

tdwi
Transforming Data With Intelligence™

JUNE 2018

TDWI CHECKLIST REPORT

Five Data Management and Analytics Best Practices for Becoming Data-Driven

By Fern Halper

TABLE OF CONTENTS

tdwi

Transforming Data With Intelligence™

555 S. Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org

FOREWORD

Analytics has hit the mainstream. The vast majority of organizations believe in the power of data and analytics to drive insight. Yet, there is a difference between using data to glean insights and analyzing data to drive decisions and actions—that is, to becoming data-driven. At TDWI, we see that organizations are at various stages of maturity when it comes to achieving this goal. For instance, in a recent TDWI Best Practices Report, only a third of respondents stated that they are data-driven.[1]

There are organizational and technology components critical for a business to succeed in becoming data-driven. On the organizational side, a key component to succeeding with data and analytics is to create a culture that supports these efforts. Companies that succeed are typically goal-driven, transparent, empowering, and collaborative. They have strong leadership that believes in data and they are governance oriented. On the technology side, the company takes steps to ensure sound data quality and has operationalized analytics to take action. Data-driven organizations often have an integrated analytics and data management strategy that spans the entire analytics life cycle from problem identification to data access and manipulation through analytics development, deployment, and monitoring.

It can be difficult to create an organization that thrives on data and analytics. It involves breaking down silos and getting people to see eye to eye, as well as building trust and collaborative team structures. It also involves enabling more people to access and work with data, which requires both skills and tools that support different personas such as data engineers, DevOps, data scientists, business analysts, and business users. This TDWI Checklist Report discusses best practices to build a program and an infrastructure for becoming data-driven.

[1] See the *2017 TDWI Best Practices Report: What It Takes to Be Data-Driven*, online at tdwi.org/bpreports.

NUMBER ONE

BUILD RELATIONSHIPS TO SUPPORT COLLABORATION

It is impossible for an organization to become data-driven if team members don't collaborate. In fact, TDWI research often cites collaboration between groups as a critical ingredient to analytics success. In addition, collaboration is linked to improved market performance and innovation.[2]

Collaboration can be hard, though. We often hear from business groups that IT has too many complex processes and doesn't understand that the business needs results quickly. IT feels that the business doesn't understand its priorities in terms of data management. Politics often enter the picture.

The key is for these groups to appreciate each other's perspective in order to move forward effectively. Of course, having executive support can help because executives set the tone and vision for the organization. However, it is important for other data and business leaders to come together as well. That means sitting down and communicating. A typical comment we hear from organizations looking to become data-driven and collaborative is, "Once IT understood what we were trying to do, it became easier to work together."

That is not to say that groups don't have individual roles—they do. In fact, there are numerous roles that come together as part of a team to make sure that analytics is successful. These roles include IT and the business as broad categories, along with more-defined roles in each category such as architects, data scientists, business analysts, and business sponsors, to name a few. Defining roles and responsibilities is important; a study by Harvard Business School found that collaboration improves when the roles of individual team members are clearly defined and understood.[3] A more recent study performed by Google (as an extension of its Project Aristotle), determined the same thing.[4]

At an organizational level, role definitions include:

- **IT/Architecture owns the data strategy.** Previous TDWI research indicates that in an analytically mature organization, IT owns data management. That includes creating the enterprise data management strategy, increasing data sharing, dealing with metadata, and owning data quality. In this way, IT can help ensure data integrity to support business decisions.
- **The business is responsible for aligning projects with organizational objectives.** Although collaboration between business and IT is essential to ensure data exists to answer the questions, the business typically owns the decisions and performs the analysis. Business users understand the kinds of questions they want answered and should therefore identify

analytics objectives. This includes identifying metrics the organization should measure from the results of its analytics efforts to determine whether it achieved its goals.

- **Roles within these functions.** There are numerous roles within the business and IT functional areas. These include business sponsor, business user, business requirements analyst, data architect, data steward, data analyst (develops data models), business analyst (performs analytics), data scientists, data engineers, DevOps, and many more, with new roles becoming the norm.

TDWI research indicates that data and analytics professionals typically fill several roles.[5] For instance, a BI director might also identify as performing data architecture or data quality roles. A data analyst might also act as a project manager and a business requirements analyst. Each role must be well-defined and people must understand their roles, tasks, and deliverables so there is clarity, turf wars are avoided, and people can perform their specific functions independently and collaborate with other functions as required.[6]

Of course, it is important for both business and IT to identify new roles and provide career paths for team members. This can include certifications to recognize expertise and deliver ongoing success.

[2] Jamrog, Jay. Purposeful Collaboration: The Essential Components of Collaborative Cultures, https://www.i4cp.com/webinar-portfolios/purposeful-collaboration-the-essential-components-of-collaborative-cultures. Accessed March 18, 2018.

[3] Erickson, Tamara J., and Lynda Gratton. "8 Ways to Build Collaborative Teams," *Harvard Business Review.* https://hbr.org/2007/11/eight-ways-to-build-collaborative-teams. Accessed March 8, 2018.

[4] Rozovsky, Julia. "The Five Ways to a Successful Google Team," https://rework.withgoogle.com/blog/five-keys-to-a-successful-google-team/. Accessed May 1, 2018. See also Duhigg, Charles. "What Google Learned in its Quest to Build the Perfect Team." *New York Times Magazine*, February 25, 2016.

[5] For example, such combined roles can be found in the *2018 TDWI Salary, Roles, and Responsibilities Report*, available to members at https://tdwi.org/research/list/salary-roles-and-responsibilities.aspx.

[6] This is described in more detail in the article, "8 Ways to Build Collaborative Teams" (see footnote 3).

NUMBER TWO

MAKE DATA ACCESSIBLE AND TRUSTWORTHY

If organizations are going to break down departmental barriers and make data a corporate asset, data accessibility and quality are critical. Although many will need access to data, different personas will have different relationships with it. For example, a data engineer might be responsible for assembling data and transforming it for analysis. The data scientist might build a predictive model using it. A best practice is to unify the data in some way and to utilize common vocabularies so all personas are on the same page.

At TDWI, we are seeing organizations integrate and access data across a multiplatform environment for analytics. This is becoming the norm, as organizations need to analyze "new" forms of data (e.g., text, sensor, image, streaming) as well as traditional forms. Organizations are making use of platforms including the warehouse, Hadoop, streaming platforms, and data lakes, both on premises and in the cloud, as part of the move to a multiplatform architecture. Therefore, the key is to unify the data for analysis.

Some important factors here include:

- **A data integration and pipeline environment.** As organizations need to access data from disparate systems for analysis, they need a coherent data integration environment. Some of the data might be structured data from a data warehouse, other data might be in a Hadoop data lake in the cloud. Tools that enable federated access to various data stores and the ability to join data across sources for analysis can be important so users don't have to physically move the data to analyze it.

  Additionally, as organizations use disparate kinds of data from multiple sources, they can benefit from tools that utilize a point-and-click interface to build a complex workflow for processing data from source systems to target systems. Vendors now provide visual design tools that support advanced specifications such as conditional logic (IF/THEN) or parallel jobs. Some of these tools support both ETL (extract, transform, and load) and ELT (extract, load, and transform). The pipelines can be scheduled to run at a specific time and may include many transformation capabilities out of the box to help make it easier to build the workflow.

- **Metadata management capabilities.** Metadata management is also critical. Metadata is data about data. It includes information such as file size, author, creation date, database column structure, table definitions, and security levels. It also includes information about data usage and interpretation. As data increases in size and scope, it is important to understand what is included in that data. Existing information about data should be reused. Metadata should be captured and coordinated and governed across all downstream BI and analytics tools where data is interpreted and consumed. Some vendors enable metadata to be propagated from source to target systems. Some provide tooling that enables metadata capture throughout the data integration and transformation process.

- **Data quality tools.** Data quality is key to trusted data and to efforts for becoming a data-driven organization. Data quality tools can profile data to identify features such as data

accuracy, completeness, and ambiguity. Data quality rules can be reused against new data. As discussed further in Number Three, some tools have advanced analytics capabilities built in to meet data quality objectives. For example, some tools use machine learning to automatically determine whether data looks reasonable or to perform tasks such as address mapping by teaching the software what addresses or people are the same.

NUMBER THREE

PROVIDE TOOLING TO HELP THE BUSINESS WORK WITH DATA

Many groups across the organization want to analyze data and take action on it—a key component of becoming data-driven—including marketing, finance, operations, and HR. Vendors have been working hard over the past few years to provide easy-to-use self-service tools to help business teams with such tasks as data preparation and analysis. Many of these tools have advanced analytics "smarts" (such as machine learning) built into them and are designed to work across the analytics life cycle, from data collection and profiling to monitoring advanced analytics models in production.

Some of the important principles of this new class of tools include automation, reusability, and explainability:

- **Automation.** Automation, which involves reducing human intervention, comes in many flavors. For instance, because data scientists and statisticians (the people who build predictive models) are in short supply, vendors are providing tools that automate the model-building process. For the business analyst persona, for example, the analyst simply specifies the target (or outcome) variable of interest and the software creates the best model using the attributes provided. Some data preparation tools are making it easier to find, collect, integrate, profile, and transform data. Some tools utilize machine learning and natural language processing to understand semantics and accelerate data matching.

  Because models get stale, vendors are also building automation into the model management process. Some tools allow a model builder to specify rules for the system to trigger alerts when the model is degrading and needs to be retrained. Other tools go further and perform automated detection of model degradation based on lift or some other parameter.

- **Reusability.** Another aspect of automation is reusing what has already been created for data management and analytics. For instance, workflows are often created using drag-and-drop interfaces to assemble data pipelines from source to target. That same workflow can be saved and embedded into an analytics workflow to create a predictive model. Many vendors provide the capability to save and reuse workflows. Some provide scheduling capabilities, as well.

- **Explainability.** As tools become easier to use, those using the tools must understand what they are doing. For example, a business user building a predictive model with an automated tool must understand what the output means. To this end, tools often include explanations of what they have done. For instance, an analytics tool using neural network technology can include with the results an explanation of how the results were derived.

NUMBER FOUR

CONSIDER A COHESIVE PLATFORM THAT SUPPORTS COLLABORATION AND ANALYTICS

TDWI research indicates that an analytics platform is a top priority for organizations looking to become more analytically sophisticated and data-driven.[7] Because there are numerous roles that contribute to the overall analytics effort, vendors are beginning to create platforms that support multiple personas. For instance, it would not make sense to purchase a data science workbench solution that primarily supports those who code in python if the average user is a business analyst. Business analysts might like a visual user interface where they can construct workflows.

As an organization matures in its analytics journey, more users will become involved in the process. A platform that supports multiple roles in a common interface, with a unified data infrastructure, is worth considering. These platforms have interfaces that support how various people perform their jobs; the platforms also enable collaboration. Below are a few examples of how this works.

- A business analyst is building a predictive model using a tool with a GUI for workflow development. While developing the model, the analyst might collaborate, via a discussion space on the platform, with the data scientist to ask questions about certain features. Additionally, the company policy is that any models built by business analysts must go through a control testing process with a data scientist to make sure the model is "good enough" to go into production. The business analyst can simply share the model with the data scientist on the same platform that utilizes the same data.

- A data scientist builds a model using an open source tool that is part of a data science workbench, which might include a notebook environment (i.e., a live, interactive programming environment). After the data scientist tests and validates the model, it is versioned and the metadata is captured. The data

scientist can then notify the DevOps team that the model is ready for production. The DevOps team—using the same platform—can view that model and use their preferred technique (e.g., APIs, PMML, etc.) to move it into production. They can continue to monitor the model with the platform's tools.

Using a platform for data and analytics collaboration makes sense—first, because there is a common data architecture underpinning the analytics tools, and second, because the platform enables different personas to share data and analyses.

[7] See the *2018 TDWI Best Practices Report: Practical Predictive Analytics*, online at tdwi.org/bpreports.

NUMBER FIVE

UTILIZE MODERN GOVERNANCE TECHNOLOGIES AND PRACTICES

Governance rules and policies set out how an organization protects and manages its data and analytics. This involves a number of factors including security, quality/accountability, access, and processes. It is important for organizations to certify and manage relevant input data sets and govern information outputs to ensure accountability, consistent understanding, and interpretation as well as to derive value from information assets. A recent TDWI survey indicates that most organizations do not do a good job of governing their data. In fact, about a third of organizations don't govern their data, at all.[8] Many organizations focus on security and privacy rules as opposed to building out more robust data governance practices. Yet, data governance is critical for an organization to trust data and become data-driven.

Additionally, our research shows that fewer than 20 percent of organizations do any sort of analytics governance. Analytics governance involves processes such as vetting and monitoring models in production. These are important tasks for the data-driven organization because it is important to monitor analytics that are operationalized and used to take action. Decisions based on poor analytics can have a negative impact on an organization.

Governance is an evolving practice. As more people become involved in analysis and as organizations deal with big data, the cloud, and other new technologies (such as stream mining), practices need to evolve. For instance, as more people analyze data, there is a greater need for consistent vocabularies and flexible access to data. Similarly, as new kinds of data come into the organization, an enterprise must determine how to deal with them. It is equally important for organizations to have policies for implementing predictive models.

On the technology front, vendors are adding features to their software to help with data and analytics governance. Some important features include:

- **Data catalogs, glossaries, and dictionaries.** The enterprise data catalog is an inventory of data assets (including metadata) that enables organizations to become data-driven by helping users understand and build trust in data. Likewise, data glossaries help users understand common vocabularies and business terms, which are important when analyzing data. Newer tools automate some of the building steps and the procedures for keeping catalogs up to date, as well as discovering metadata from existing data sets to learn details about data. The tools can also tag data according to higher-level business definitions and rules, and locate and use existing documentation.
- **Data lineage.** Data lineage enables users to trust their BI and analytics. Using data lineage with metadata, organizations can understand where the data originated and track how it was changed and transformed. This could include how the impact of changes to one data element might affect other systems.
- **Model management.** Once a predictive model is built, it can get stale and degrade over time, so continual tracking is an important part of the analytics governance process. Some vendor solutions automate monitoring; they can schedule updates for model building using a champion/challenger approach to make sure the model is still current. Other tools provide automated alerts when a model is degrading.

Involving IT and business stakeholders is key to robust governance. This governing group needs to embrace flexibility and agility. The same kind of human factors mentioned previously are important to help build relationships among data owners, data stewards, and others involved in governance.[9]

The approach organizations take to governance can vary based on maturity level and objectives. Some newer models being discussed in the market move beyond the traditional governance council model. These include:

- **Agile governance.** In this structure, governance begins with the planning and development team. These stakeholders—functional, operational, business sponsors, subject matter experts—understand governance needs and share them as constraints, expectations, and requirements for the development team. The development team is then responsible for including governance needs in the development process and communicating with stakeholders to fully understand governance goals.
- **Embedded governance.** In this model, the idea is to flatten the governance organization by integrating governance into

business functions and projects. For instance, governance team members such as data stewards are embedded in a business team.

- **Crowdsourced governance.** Crowdsourcing is the practice of engaging a large number of people via the Internet for a common goal. This (somewhat controversial) concept was introduced a few years ago for data governance. The idea is to engage some number of people (e.g., the data community) to report data quality problems, propose metrics, and provide input or even vote on standardized terms. Of course, whether or not a company decides to go this route depends on its culture and what it is trying to accomplish with its governance program. An operating model is still needed for this to work.

[8] From the *2017 TDWI Best Practices Report: What It Takes to Be Data-Driven*, online at tdwi.org/bpreports.

[9] Training in data governance can be found online, for example TDWI's Data Governance Innovations online at https://tdwi.org/events/onsite-education/onsite/sessions/data-management/biz-all-tdwi-data-governance-innovations.aspx

FINAL THOUGHTS

Organizations are facing new challenges when it comes to data and analytics. Data is increasingly diverse. Analytics is becoming more sophisticated. New users want access to data and to perform analysis. To become data-driven, a business needs to both embrace and plan for this reality.

Becoming data-driven involves people, processes, and technologies. People includes building relationships that create trust. Communication is critical here as is understanding other points of view. Organizational leadership can help drive collaboration via training and policies that support relationship building.

Processes mean collaboration that can include a framework for how people can work together, including specifying roles and responsibilities so people know how they are contributing to the overall goal. Collaboration also includes a governance structure to ensure that trusted data is used for analytics put into production.

Technology can help organizations become data-driven, too. Vendors are stepping up to the plate to offer newer features that make it easier for both business and IT to access and analyze the increasing scope and scale of data . This includes better tools for assembling data pipelines and analytics workflows. It also includes analytics platforms that support multiple personas with the same experience and integrated data environment, and tooling that supports automation by embedding sophisticated analytics into the software. These technologies make the analytics life cycle more streamlined and efficient.

It is important to talk to vendors about what they offer. Dig deep with them to understand if the software will meet your needs now and into the future, then pick a vendor you can trust.

ABOUT THE AUTHOR

Fern Halper, Ph.D., is vice president and senior director of TDWI Research for advanced analytics. She is well known in the analytics community having been published hundreds of times on data mining and information technology over the past 20 years. Halper is also coauthor of several Dummies books on cloud computing and big data. She focuses on advanced analytics, including predictive analytics, text and social media analysis, machine learning, AI, cognitive computing and big data analytics approaches. She has been a partner at industry analyst firm Hurwitz & Associates and a lead data analyst for Bell Labs. Her Ph.D. is from Texas A&M University. You can reach her by email (fhalper@tdwi.org), on Twitter (twitter.com/fhalper), and on LinkedIn (linkedin.com/in/fbhalper).

ABOUT TDWI RESEARCH

TDWI Research provides research and advice for BI professionals worldwide. TDWI Research focuses exclusively on analytics and data management issues and teams up with industry practitioners to deliver both broad and deep understanding of the business and technical issues surrounding the deployment of business intelligence and data management solutions. TDWI Research offers reports, commentary, and inquiry services via a worldwide membership program and provides custom research, benchmarking, and strategic planning services to user and vendor organizations.

ABOUT TDWI CHECKLIST REPORTS

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, analytics, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.

Becoming a data-driven organization requires proof that decisions based on data are better than those that aren't. So how do you prove it? By defining priorities for improvement and innovation, then using analytics to explore options, make evidence-based decisions and implement the best course of action.

Often that's achieved by embedding SAS® Analytics directly into business operations and processes. SAS customers reap the benefits of increased productivity, improved operational efficiency, cost savings and top-line growth by analyzing data and driving measurable value in their area of specialty.

Some organizations have a wide variety of data and analytics skill sets. Serving wide-ranging needs, SAS provides interfaces that meet the skills of knowledge workers. Data from a variety of sources – whether Excel, RDBS, Hadoop or others on-site, in the cloud or both – is readily available for key workers to make the right decision, using the right data at the right time.

Successful organizations recognize that analytics models are essential corporate assets that yield better customer relationships, improved operations, increased revenues and reduced risks.

But creating the best models possible requires managing the complexities of the three parts of the analytics life cycle – data, discovery and deployment. At SAS, we've developed a proven, iterative analytics life cycle to guide you step by step through the process of going from data to decision.

As mentioned in the TDWI checklist report Five Best Practices for Being Data Driven, organizations must do the following:

1. Build relationships to support collaboration.
2. Make data accessible and trustworthy.
3. Provide tools to help business work with data.
4. Work with a single platform that supports collaboration and analytics.
5. Utilize governance.

How SAS® Meets the Five Best Practices

Build relationships using a single platform to support collaboration and analytics.
The SAS Platform is an analytics platform engineered to generate insight from data in any computing environment and use those insights to drive business actions. The platform supports every phase of the analytical life cycle – from data to discovery to deployment.

Make data accessible and trustworthy.
You can't make decisions based on data unless you're certain it's trustworthy and accessible. The SAS Platform is built on more than 40 years of best practices that have gone through extreme vetting for usability, reliability and functionality. Since data is one of the foundational pieces of the analytics life cycle, data access and data quality capabilities are critical to the decision making process.

Provide tools to help business work with data.
Analytics tools must support various roles, and SAS is designed to help business users easily integrate, browse and clean data in a sharable environment with a self-service approach in mind. The goal is to get the right data in the hands of the users who need it as quickly as possible, and give them the ability to adjust the data for faster insights.

Work with a single platform that supports collaboration and analytics.
Our single platform helps build collaboration between departments and across skill sets. By having a single platform for data, analytics and reporting, all parts of the organization are empowered to work together toward the common goal of being data driven.

Utilize governance
Addressing the entire analytics life cycle as an integrated process, SAS provides common components that communicate with one another, ensuring the transitions between phases of the analytics life cycle are seamless and fast. This includes authorization, security, lineage, governance, common vocabularies, search and centralized management.

Applying results deployed in mobile dashboards, as alerts, as new data elements in data lakes, or via transactional systems, SAS shares knowledge with others to act upon. You can centrally monitor and control the management of automated data-driven actions embedded in transaction systems, operations and event streams, ensuring the health of data-driven results.

SAS provides scalable, easy to use, business-designed software based on more than four decades spent addressing the diverse needs of enterprise analytics. Because the world's top organizations embrace data-driven insights and decisions, they count on SAS to produce results that make a difference.

www.sas.com/platform