research

CHECKLIST REPORT

2018

Six Keys to Maximizing Big Data Benefits and Project Success

How you can deliver business value and avoid common pitfalls in big data integration and analytics

By David Stodder

Sponsored by:

HITACHI
Inspire the Next

tdwi
Transforming Data
With Intelligence™

SEPTEMBER 2018

TDWI CHECKLIST REPORT

Six Keys to Maximizing Big Data Benefits and Project Success

How you can deliver business value and avoid common pitfalls in big data integration and analytics

By David Stodder

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

TABLE OF CONTENTS

Product and company names mentioned herein may be trademarks and/or registered trademarks of their respective companies. Inclusion of a vendor, product, or service in TDWI research does not constitute an endorsement by TDWI or its management. Sponsorship of a publication should not be construed as an endorsement of the sponsor organization or validation of its claims.

FOREWORD

Big data is about changing the status quo for established organizations and fueling the growth of new and disruptive businesses. Big data projects focus on enabling analysis of and interaction with new types and combinations of data at a far greater scale than has been possible with traditional enterprise business intelligence (BI) and data warehousing systems.

Both established and new organizations need systems that can support not just standard BI but also advanced analytics and cutting-edge information-driven applications that use artificial intelligence techniques such as machine learning.

Many organizations are modernizing their data architectures by surrounding existing systems with technologies based on Apache Hadoop and Apache Spark ecosystem open source projects and cloud-based data storage and processing—offloading work to more economical and higher-performing big data platforms.

Of course, some organizations are not simply augmenting existing systems—they are replacing them with big data lakes built with Hadoop and Spark ecosystem technologies and/or cloud-based data platforms. They need integrated, end-to-end technology solutions that help them reduce manual effort and custom coding so they can scale to handle more users, data-intensive requirements, and stringent performance expectations.

Many new and disruptive businesses have never had traditional BI and data warehousing systems; they began their journey with big data platforms. They need solutions and practices that enable them to govern and manage big data projects effectively and derive higher value from them, including through support of BI-style workloads and data transformations needed for business analysis and reporting.

In each scenario, organizations face challenges in avoiding pitfalls and overcoming hazards that can frustrate visions of using big data to achieve 360-degree customer insights, smarter operations, more effective risk management, and other critical goals.

Fortunately, technology solutions are available today that build on lessons learned from first-generation big data projects to help succeeding generations avoid delays, poor integration, significant manual work, and errors that have prevented organizations from fulfilling their ambitions with big data and analytics. Better integrated and easier-to-use technologies plus guidance provided by blueprints and templates can enable organizations to move faster to realize the benefits of big data systems for a wider range of strategic and operational business decisions.

This TDWI checklist discusses six important issues that organizations should address to start big data projects off right and then manage them to achieve objectives faster and with less difficulty. By addressing these issues, organizations will be in a better position to accomplish extraordinary results that change the status quo.

☑ NUMBER ONE

REDUCE TIME TO VALUE BY CLARIFYING OBJECTIVES AND DELIVERING BENEFITS SOONER

When it comes to discovering and applying data insights, sooner is always better. A big data project can contribute most to achieving business goals when leaders shorten the time to value, that is, the time interval from a project's inception based on perceived business needs to its ultimate delivery of anticipated value. TDWI research finds that two common obstacles to reducing time to value are poor project definition and scoping and a shortage of skilled personnel. Solutions can help organizations address both of these issues.

The first step is to carefully define the business case and articulate how the project will contribute to a key objective, such as increasing sales and market share, reducing costs, or generating new products and services for partners or customers. Data exploration can play an important role in defining a business case. For example, an organization might want to know what's behind an unanticipated trend in customer buying preferences. A new organization seeking to be a disruptive force in a market needs data insights to help it clearly understand customer needs and where established players are vulnerable. Business leaders in each case will want to look for patterns and correlations across data sources to test theories.

Easy-to-use, visual data discovery tools that are well-integrated with supporting data extraction, transformation, and loading (ETL) are valuable at this stage for enabling nontechnical users to interact with data and set up data pipelines for analytics and machine learning. This puts less pressure on the organization to hire skilled specialists. In addition, with such tools, organizations can help their current data scientists and analysts spend less of their time on manual efforts to find and prepare data. Business users, analysts, and data engineering team members can use visual data interaction solutions to explore whether there's justification for a substantial big data project, such as a new, data-intensive application that monetizes data by selling analytics services to business partners.

Once a project gets going, the second step is to deliver incremental benefits. Too often with traditional BI and data warehousing, project development takes months, if not years; by the time systems are finished based on the original specifications, user requirements have changed. Organizations should consider using agile methods that support strong teamwork between business leaders, analysts, data scientists, developers, and data engineers and shorter, incremental development cycles. Users can then work with intermediate deliverables, perhaps do more data exploration, provide feedback to developers, and then test the next iterations. Again, easy-to-use tools that integrate data discovery, ETL, analysis, and visualization can play a vital role in allowing users to sharpen data insights with each iteration and refine project objectives.

Over time, through testing and collaboration, organizations can refine the project's scope. The team will have the agility to evolve the project to fit changing business requirements. Self-service solutions can help organizations avoid having to wait to realize value until they are able to hire enough specialists to build complete applications and systems.

☑ NUMBER TWO

USE BLUEPRINTS AND TEMPLATES TO SIMPLIFY PLANNING AND FOCUS ON VALUE

Excited by the potential of big data analytics, organizations will often launch into implementing technology without a good plan for how they will assemble resources and orchestrate them to deliver value.

For example, an organization may sense that if it could integrate customer data from transactional, behavioral, social media, and support engagement sources—impossible with its traditional BI and data warehousing systems—it could gain greater insight into customer preferences, enabling it to sharpen sales and marketing campaigns. The team will then immediately start pouring big data into an on-premises data lake or cloud data storage platform. Unfortunately, because the organization did not set out a plan for how it would integrate and organize the data to make it easier to uncover its value, business leaders become frustrated and question the project's merits.

Big data projects are often complex due to the variety and volume of data and the sophistication of project plans and intended business outcomes. Templates and blueprints are therefore critical for big data projects because they enable organizations to start with something more than just an empty canvas.

Organizations can use templates and blueprints to see how projects similar to theirs have been done before, how different pieces of the whole fit together, and how they may need to customize the plans to fit their unique requirements. Templates and blueprints also support reuse of proven processes and routines, which can save organizations time and money by avoiding reinventing wheels unnecessarily.

Templates and blueprints are available from solution providers, consulting firms, vertical industry associations, and some Apache open source projects. Organizations should use them first to get the big picture: a lay of the land for how different phases of projects should fit together, such as the steps for properly filling a data lake. Organizations can then focus on more specific use cases, for example, enabling integrated customer 360-degree views and determining how these will be used by marketing, sales, and customer support. They can use the templates and blueprints to align business and technology requirements for achieving deliverables. This documentation is useful for communicating with the business and IT stakeholders responsible for funding projects and assembling resources.

☑ NUMBER THREE

OPTIMIZE USE OF TECHNOLOGY PLATFORMS FOR DATA INGESTION AND TRANSFORMATION

Big data lakes, whether on premises or in the cloud, can easily turn into useless big data dumping grounds if organizations do not think carefully about how they fill them so they can maximize their value. Data lakes, which usually contain a mixture of diverse data types in their raw, natural format, have proven to be a convenient place to put data for analytics and for the execution of machine learning algorithms. However, to support these activities, a data lake must be managed as more than just a storage location.

Many data lakes are built on Hadoop or Spark, which means that organizations need to keep up with technology developments in these Apache open source ecosystems. The choice of technologies should be guided by what the organization needs to accomplish with its big data platforms. As noted in Number Two, a template or blueprint can help organizations see how technologies and platforms need to fit together depending on the purpose.

Once organizations determine where the data will come from—for example, from log files, Web servers, contact center systems, transaction systems, or machine data sources—they can use the horizontally scalable computer processing power of Hadoop or Spark massively parallel processing (MPP) clusters to collect raw data, ingest it into the data lake, and transform it into a usable form. Most organizations will want to avoid running ETL jobs against production database sources themselves because these systems' primary responsibilities are to record the data at a high performance level.

Organizations should examine how they can offload ETL and other resource-intensive data preparation tasks from expensive data warehouses and specialized data integration systems by pushing them down to less expensive big data systems, either on premises or in the cloud. Tools and solutions in the marketplace can automate and standardize data ingestion, collection, and ETL processes for better speed and consistency. The ability to push down processing to big data platforms becomes essential as data volume and complexity rise.

Although for some data science and analytics projects, raw, "dirty" data is exactly what users want, clean and consistent data is preferred for the majority of projects. Many machine learning algorithms will produce more useful results if run against better quality data. Organizations should use blueprints and templates to orchestrate ETL and other data preparation processes and set up how they will use the power of big data systems during ingestion, collection, and integration. This will give organizations a cleaner data lake that is more immediately useful.

☑ NUMBER FOUR

MAKE BIG DATA PLATFORMS ACTIVE REFINEMENT CENTERS FOR UNLOCKING POTENTIAL VALUE

As organizations ingest and collect the data into the data lake or other big data repository on premises or in the cloud, the critical activity is to refine the data as soon as possible so that it is suitable for project requirements. Refinement, which includes data preparation steps for enrichment, transformation, and improvement of data quality and consistency, has traditionally involved much manual work, slowing down progress in realizing value from the data. Organizations should evaluate tools and solutions in the marketplace that can apply smart automation to data refinement and preparation. Self-service capabilities are evolving to make it easier for users who have less expertise with data to use visual interfaces rather than programming to direct these processes.

An important technology system that can support data refinement and preparation, as well as data integration and other tasks, is the data catalog, glossary, or metadata repository. Ideally, such a system provides accurate, up-to-date, and comprehensive data definitions and information about how different data sets originating from different sources are related. A data catalog can be a kind of Rosetta stone that enables users, developers, and data scientists to find and learn about data—and information professionals to properly organize, integrate, and curate data for users. Enabling access to this knowledge base during each step for data refinement and preparation will improve overall speed, completeness, and accuracy.

Defining data and articulating how different data sets are related require the input of subject matter experts who know the data best. However, as with data refinement and preparation, available data cataloging technology solutions can take it from there, automating development and maintenance of data catalogs, glossaries, and metadata repositories. Some solutions employ machine learning to learn data definitions and relationships in massive data volumes and supply this knowledge to users through the data catalog.

Data refinement, preparation, and cataloging processes are vital to making the big data lake an agile resource that more rapidly serves a variety of needs, from BI dashboards and reporting to advanced predictive analytics and machine learning. Users in marketing, for example, could tailor their refinement steps to help them spot data relationships for customer segmentation and personalization. A finance department could create refinement steps that improve forecasting by looking at different variables and the probability of how they might impact business performance. In this way, organizations can make the big data lake an active resource for unlocking value from data rather than merely a place to store it.

☑ NUMBER FIVE

TAKE ADVANTAGE OF CENTRALIZING DIVERSE DATA TO SUPPORT INNOVATIVE APPLICATIONS

Data is the lifeblood of innovative applications and services. Organizations that are seeking to disrupt markets and burst ahead of competitors need applications and services that offer more than what traditional systems provide, not only for internal users but also for external partners and customers. Centralizing data access outside the limitations of traditional systems is important to building and running new, data-intensive applications and services. Big data projects need to provide the virtues of centralized data access for a universe of far more diverse data.

With traditional data warehousing, data centralization comes at a cost: through the data warehouse, the organization has access only to a limited set of structured, formally defined and organized data. A comprehensive data lake can enable expansion of the data architecture to centralize access to and interaction with diverse data that is structured, semistructured, and unstructured. However, the challenge is to meet each workload's requirements with the right underlying infrastructure because not all workloads are the same.

For example, some organizations will choose a classic data warehousing system, either on premises or in the cloud, because it is fit for the purposes of traditional BI reporting and dashboard-based analytics. The role of the data lake is then to augment the traditional system by storing, managing, and providing access to other types of data. Other organizations will choose to make the data lake the central platform for their data architecture. In this case, the demands of applications and services need to be met through providing direct query access to data lakes and implementing technology layers that handle refinement and preparation. Whether the role of a big data platform is to augment or replace the data warehouse, organizations should consider how cloud-based systems can support rapid scaling.

Use cases for data-intensive applications and services tend to be dynamic; for example, an organization will need them to support a product launch or to analyze a specific threat or opportunity. Cloud-based resources can scale on demand based on immediate business requirements rather than depending on the availability of fixed IT assets on premises. With critical data located on premises and in the cloud, many organizations need to develop a multiplatform data architecture that enables users to view and access relevant data across platforms and allows IT personnel to manage and govern the data wherever it is physically located.

Centralizing data across a multiplatform architecture will be critical for many organizations to drive development of data-intensive applications. These applications need to access multiple data sources to feed embedded predictive models, machine learning, and other AI to crunch through diverse volumes of data fast enough to drive application and business processes. These new applications will enable organizations to monetize their data, for example, by developing services for business partners and customers that deliver timely insights about customer churn, factors influencing buying behavior across channels, predictive insights into potential problems in machinery or other facilities, and more.

To prepare for data-intensive applications, organizations need to plan for appropriate scale, availability, and speed of data refinement, transformation, and preparation. Organizations should evaluate data management tools that enable them to decide where to allocate BI, analytics, or machine learning workloads—whether to the cloud because that's where important data is located or to meet immediate business needs; to a Hadoop or Spark cluster because the application needs scalable processing power; or to the traditional data warehouse because it demands highly structured and predictable data.

☑ NUMBER SIX

MAKE DATA GOVERNANCE AND STEWARDSHIP PRIORITIES, NOT AFTERTHOUGHTS

As big data platforms expand and spawn numerous data-intensive applications, organizations have to ensure that they are governing data appropriately and provisioning applications with quality data. Data governance typically involves defining rules and policies for protecting sensitive assets, such as customers' personally identifiable information, and setting up reporting and auditing functions for demonstrating adherence to internal rules and policies as well as industry or governmental regulations, such as HIPAA and the European Union's GDPR.

Governance is not just about protecting stored data; it is about how the data is used across the organization and externally. For this reason, governance intersects with data stewardship, which is focused on improving data quality, analytics modeling quality, selection of new data assets, and ensuring user satisfaction and productivity with data. Just as they do with governance, organizations can set up quality standards as part of data stewardship to provide metrics for data provisioning and use.

TDWI research finds that too often, organizations don't address governance or stewardship until late in the process when projects go into production. Often this is when the organization's legal team is raising questions about regulatory adherence or concerns about potential exposure of sensitive data. It is wiser to address potential data governance issues at the early stages of big data projects so teams can anticipate concerns and develop analytics, machine learning, visualization, and other capabilities with awareness of governance requirements.

Historically, big data lakes have been notorious for having inadequate governance and haphazard data quality and consistency, in part because they were initially set up for data science exploration, not business-critical analytics. Fortunately, the security of data lakes is improving, particularly in the cloud. At this point, some cloud-based systems offer better protection of data for governance and security than many organizations' on-premises systems.

Nonetheless, organizations should satisfy concerns that security and governance rules and policies cover data lakes and multiplatform architectures. Organizations should set up processes for evaluating the ongoing effectiveness of governance so they can improve the rules and policies and ensure they fit the specifics of BI, analytics, and AI use cases.

Finally, organizations should expand the scope of data governance to include steps for enabling greater trust in the data and transparency in terms of data lineage—the data's provenance in terms of where the data used in BI, analytics, and AI came from, how it was transformed and by whom, and how the data is being used. The GDPR and other regulations are demanding greater visibility into how decisions (such as credit and loan approvals) are made. Data lineage is important to documenting such decision processes. Organizations should use their data catalog, glossary, or metadata repository to make it easier to track data lineage and monitor use of potentially sensitive data, no matter where it is physically located.

A FINAL WORD

We'll close this checklist by noting the importance of culture. Analytics insights that result from big data projects will frequently challenge conventional wisdom and advocate change to the status quo. In many organizations, executive leadership can be resistant to such challenges. Analytical conclusions could run counter to "the highest paid person's opinion" and risk being dismissed out of hand. Thus, it is important for organizations to think about their culture and address people and communications issues that might prove to be the biggest obstacles to realizing value from big data projects.

Leadership and communication are vital to building a healthy culture for analytics. Executives should set a receptive tone through meetings, discussion, and test-and-learn cycles for determining if the analytics is valid and useful. In this way, organizations can engage in a cycle of learning that will improve analytics model development and data management and help team members adjust focus to make sure that project definitions and deliverables are on target.

Project teams themselves need to build skills in communicating data insights to leadership, including how conclusions were derived, what they know about the data, and the relevance of insights to desired outcomes. Good communication will improve receptivity to data insights and encourage the organization to put them into action.

ABOUT OUR SPONSOR

HITACHI
Inspire the Next

hitachivantara.com

Data is your greatest asset, if you know how to use it. It reveals your path to innovation and new ways for you and the world to work. Hitachi Vantara elevates your innovation advantage with 100 years of OT and 60 years of IT experience that connects business, human, and machine data to create solutions that drive benefits for your business and society. Our unique Stairway to Value model uses machine learning and artificial intelligence to deliver tangible benefits driven by your data. We're a catalyst for intelligent innovation and we accelerate insights. Let's talk about your new vantage point.

ABOUT THE AUTHOR

David Stodder is senior director of TDWI Research for business intelligence. He focuses on providing research-based insights and best practices for organizations implementing BI, analytics, data discovery, data visualization, performance management, and related technologies and methods and has been a thought leader in the field for over two decades. Previously, he headed up his own independent firm and served as vice president and research director with Ventana Research. He was the founding chief editor of *Intelligent Enterprise* where he also served as editorial director for nine years. You can reach him by email (dstodder@tdwi.org), on Twitter (@dbstodder), and on LinkedIn (linkedin.com/in/davidstodder)

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.