research

CHECKLIST REPORT

2017

New Strategies for Visual Big Data Analytics

How organizations can apply modern data platform technologies and practices to support analytics innovation

By David Stodder

Sponsored by:

ARCADIA DATA

tdwi
Transforming Data With Intelligence™

APRIL 2017

TDWI CHECKLIST REPORT

New Strategies for Visual Big Data Analytics

How organizations can apply modern data platform technologies and practices to support analytics innovation

By David Stodder

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

FOREWORD

Excitement continues to grow about how organizations can realize the potential of big data through powerful analytics and data visualization. Executives and managers in lines of business and departmental functions want to gain more comprehensive situation awareness and better understanding of critical trends and predictive patterns through analysis of large volumes of diverse data—data that could be structured, semistructured, or unstructured—that is increasingly streaming into systems at or near real time. It is a competitive differentiator if more personnel are able to apply insights easily to daily decisions or build data-centric applications to drive new lines of business or analytics services.

Capabilities for visual big data analytics are maturing at the same time as technologies for self-service business intelligence (BI) and visual analytics. These products are democratizing data access, enabling less technical business users to get beyond "one size fits all" BI reporting and the difficulties of implementing online analytical processing (OLAP). Users can more easily personalize how they interact with data. However, most users have only been able to access a narrow slice of the data universe—that is, only what is structured and carefully prepared in the data warehouse, data marts, and data cubes. Big data beckons.

Organizations need a strategy for a modern data platform that can support users who need more than traditional BI and OLAP provide but do not have the specialized, hands-on big data and analytics skills of advanced data scientists. They need easier, highly visual, self-service data interaction capabilities for big data.

A modern data platform must cover:

- The spectrum from reporting on historical data to analytics to real-time streaming data
- Structured, semistructured, and unstructured data
- Both on-premises and cloud-based data (because big data now resides in both places)

Data access, interaction, and management must not only include traditional SQL-based tools and database management but also embrace NoSQL, search engines, event processing, and message-oriented middleware.

Much big data is stored using the Hadoop File System (HDFS) and on distributed computing platforms that support Hadoop clusters. Thus, most organizations need to examine technologies in the Hadoop ecosystem, which now includes Apache Spark, Kafka, and other critical open source programs and frameworks.

With the growing importance of cloud-based platforms as repositories of big data, organizations should also explore technologies that work with popular cloud-based data storage so that users can access the data with fewer intermediate steps.

This TDWI Checklist focuses on six key modern data platform considerations for enabling more users to benefit from big data through more easy-to-use, visual big data analytics.

NUMBER ONE

ADDRESS EMERGING REQUIREMENTS FOR BROADER USE OF BIG DATA ANALYTICS

As organizations expand data-informed decision making, old models for how knowledge workers across the spectrum interact with and analyze data are under pressure. For those who value flexibility in their choice of data sources, traditional BI and data warehousing environments are too confining. These environments are not supporting users who need to perform comprehensive, detailed analysis and visualization involving diverse and voluminous big data.

Self-service BI and analytics tools have made visual data interaction easier for nontechnical users, but most tools cannot handle the variety, volume, and velocity of big data. Users are limited to traditional BI methods—consuming extracts from a small selection of data to which they are able to apply only limited styles of analysis.

At the other end of the spectrum is data science. To mine for insights in big data and apply advanced techniques such as machine learning, many organizations have sought expert data scientists who can apply advanced skills in statistics, predictive modeling, artificial intelligence, and programming. However, expert data scientists are not easy to find and retain, particularly those who also have relevant business domain knowledge.

This leaves a growing population of knowledge workers in the middle of the spectrum, who some industry observers call "citizen data scientists." Primarily working in business functions and lines of business, these knowledge workers want to perform more advanced analytics than they are able to with self-service BI tools. However, being less technical than data scientists, they need easier to use, GUI-based tools that don't require manual, hands-on scripting, coding, and math.

An example would be managers in a marketing function who seek to make data-driven decisions and apply big data insights to personalization, segmentation, and recommendations. These knowledge workers—often responsible for both digital and conventional marketing—need to build models to help them allocate spending across channels.

These managers need to perform detailed analytics on customer behavior data, which could be stored in the cloud or on Hadoop clusters, to improve segmentation and personalization. Battling in competitive markets, they cannot wait for IT BI teams or the few data scientists to do this for them; they therefore require tools that offer self-service capabilities for working with big data.

Organizations need to recognize the emerging analytics requirements of their citizen data scientists. These personnel are often involved in projects that relate directly to the firm's ability to realize competitive advantages through serving customers and partners faster and with precision and intelligence. Organizations should evaluate products that can fill the void between self-service traditional BI and specialized data science.As organizations move forward with advanced analytics and data science, they must also refresh BI and online analytics processing (OLAP) strategies so that users have alternatives that enable them to take advantage of big data. Traditionally, BI and OLAP have required lengthy processes that rely on purpose-built technologies for the careful selection and preparation of structured data extracts, cubes, and aggregations. Processes for preparing the data can be slow and labor-intensive. Depending on the volume, quality, and complexity of the data, the processes can take days or even weeks.

To be sure, various reporting and managed data access requirements will still require moving selections of big data through such processes and into a data warehouse before users can access the data. However, organizations should examine whether different methods and technologies would serve users' needs more effectively. Traditional processes are not optimal when the underlying data and data structures vary or are changing rapidly, which is frequently the case with big data. Prebuilt cubes designed to serve OLAP needs are not constructed to meet time-sensitive data requirements and can easily become stale.

NUMBER TWO

IMPROVE BI, OLAP, AND VISUAL ANALYTICS THROUGH EASIER ACCESS TO BIG DATA

As organizations move toward more frequent, near real-time data updates, it can require significant effort to keep the virtual views and physical data extracts, cubes, and aggregations up to date. For example, users in marketing and sales must meet business demands for faster decision cycles so that they can respond promptly to changes in customer preferences.

To make decisions, users need to view customer intelligence trends and insights drawn from big data analytics in their dashboards or other visualizations. Traditional processes for BI and OLAP frustrate users by being too slow to build, update with new data, and reorient to support new types of analysis.

At this point, organizations should evaluate technologies and methods that will provide easier and faster access to new and changing big data. Technologies can now enable development of prebuilt cubes and similar structures on big data sources for common queries or standard requirements such as regulatory reporting. For the alternative use case of data discovery and exploration, technologies can provide users with direct access to big data rather than only having access through cubes, extracts, and aggregations (direct access will be discussed in the next section).

With big data, users often prefer to start with search and exploratory discovery rather than predefined reports and data cubes. Search engines, such as those built using open source Apache Solr, are becoming important to improving the range and speed of exploration of voluminous and diverse big data. Search and discovery are critical capabilities for supporting the new generation of applications built on big data platforms and organizations should ensure that they are available so that users can find relevant data faster.

Finally, organizations should evaluate technologies that enable users to produce reports and visualizations more dynamically as they discover relevant big data. Organizations should examine technologies that can automate development of templates and preconfigured views, which can improve report consistency and repeatability as well as the sharing of big data insights.

Capabilities should not limit users to reports; users should be able to perform ad hoc queries on detailed big data to satisfy questions that reports do not answer. These capabilities should enable users to develop metrics, measures, and dimensions with big data and integrate them with those that they are using in traditional BI and OLAP systems.

NUMBER THREE

EXECUTE VISUAL BIG DATA ANALYTICS PROCESSES NATIVELY WHERE DATA IS STORED

As noted in the previous section, traditional BI, OLAP, and data warehousing environments have a lot of moving parts, and shifting data between them can slow business insight.

For big data analytics processes that require faster access to a wide range and volume of data, organizations should consider doing more work natively where the data is stored rather than moving the data to intermediate systems. For big data, this usually means executing analytics natively in the Hadoop cluster where data files are stored or in cloud storage systems such as Amazon Simple Storage Service (S3) or Microsoft Azure Storage. Technologies are now available to enable users to execute queries and analytics inside Hadoop clusters and/or cloud-based storage rather than having to move the data to a different platform such as a staging area, data warehouse, data mart, or a standalone BI server.

The Hadoop ecosystem provides diverse open source and commercial technologies for native processing of fast, interactive BI queries, data discovery and exploration, and sophisticated analytics processes such as testing predictive models.

Organizations that have a data lake using Hadoop ecosystem technologies can use tools to prepare data, including transformations, in the clusters rather than moving it, or they can run analytics directly on the raw, detailed data. Considering that the data of interest could be a combination of standard transactional data, text content, social media data, images, network logs, and sensor data, the ability to keep it intact in one big data lake can be an advantage.

Among the open source frameworks and technologies useful for processing analytics natively on data in Hadoop files and clusters are Apache Hive for large batch SQL processing; Apache Impala, a low latency query/analytics engine for Hadoop; Apache Drill, a low-latency query engine useful for schema-free data exploration; and Apache Spark SQL, which takes advantage of Spark in-memory computing.

The linear scalability offered by Hadoop clusters—and similarly, the elastic scalability of cloud-based storage—can enable organizations to be agile and flexible in expanding computing power in response to immediate BI and analytics needs. Organizations can also centralize security and governance more effectively—or use security procedures already set up at the sources—if preparation and processing happen where the data is located.

Here, organizations should evaluate how to make the fullest use of the Hadoop ecosystem by running BI and analytics processes, including preparation steps, directly inside the environment. Organizations should evaluate whether technologies that enable running processes natively where the data is stored can better support their range of data interaction and analysis use cases—from reporting on historical data to visual, interactive querying and discovery to more advanced, predictive analytics.

NUMBER FOUR

DELIVER VALUE FROM REAL-TIME DATA STREAMS WITH BIG DATA ANALYTICS

Real-time insight for mainstream business objectives is no longer a fantasy. Technology advances are enabling organizations to shift from total dependence on historical data sets and batch processing to being able to monitor real-time data streams, ingest real-time data, and perform analytics on data in near or actual real time. To sense and respond intelligently for real-time customer engagement, process management, and situation awareness, organizations need insight into what is happening now, not just what has happened.

If organizations can analyze real-time data streams along with historical data, they can gain new perspectives on trends and take action sooner.

Real-time data sources run the gamut from Internet of Things (IoT) sensors to financial transactions, social media activity, Web clicks, and other online customer behavior. Mobile devices, aircraft, and connected cars and trucks are becoming important sources of real-time data. Organizations can combine real-time data streams with maps to enable current location analytics.

Other types of visualizations can help users look at real-time data streams and historical data together to see correlations and spot anomalies or variances from predicted patterns. Such insights are valuable across many industries, including logistics, for maintenance and driver safety; healthcare, for monitoring population health and emergency care; and insurance, for risk management and fraud detection. Internally, IT can apply streaming analytics to cybersecurity for faster detection of network and system attacks and to monitor performance and availability.

Organizations should evaluate technologies for ingesting, managing, and analyzing real-time data streams and integrating information from them with views of historical data. An emerging industry framework for bringing together real-time and historical data is the lambda architecture. It sets out how batch, streaming, and serving layers fit together.

Organizations should also evaluate the Hadoop ecosystem, which is moving beyond its initial batch processing orientation to support use of real-time data. Relevant Apache open source projects include Storm, Kafka, Spark Streaming, Flume, Samza, and Kudu. Leading commercial products integrate with and support these technologies to provide fuller solutions.

Choosing the right technologies to implement depends on the use case. In some instances, organizations just need to ingest streams and persist event data on storage systems such as HDFS, Kudu, or Amazon S3 for historical analysis. In other cases, they may want to monitor streams, run predictive models, and set up alerts so they can respond proactively as events are happening. Flume is often used for simple ingesting of data streams. Storm event processing can complement Hadoop and facilitate running analytics on a continuous stream of data. Kafka provides a scalable publish-and-subscribe messaging system that employs compression to optimize performance and mirroring for higher availability and scalability.

The Spark Streaming component can fit into the lambda architecture to support use of Kafka, Flume, and other standards. Organizations can also apply Spark machine learning components on big data streams. Kudu, one of the newest Apache open source technologies, has a columnar storage system that can support time-series analytics, fast analytics on machine data, and interactive BI on changing data.

Organizations should assess how streaming analytics and other uses of real-time data could contribute to key objectives such as improving customer engagement, reducing downtime, and increasing process efficiency. Organizations can gain initial success with smaller projects that have clearly defined goals. These will offer validation to support larger and more complex projects.

NUMBER FIVE

USE BIG DATA ANALYTICS TO POWER NEW DATA-DRIVEN APPLICATIONS

The revolution in big data analytics is changing application development. Organizations need to update their application strategies to take advantage of big data analytics. They should prepare their data architecture for new types of applications that rely on continuous, highly available big data analytics.

In recent years, social media and Internet companies have pioneered the development of data-driven applications, which generate and/or use huge volumes and varieties of event, content, and consumer behavior data. The pioneers have applied big data analytics, including machine learning, to understand and predict future behavior. Real-time analytics determines how data-driven applications behave in response to consumer activity.

Firms in other industries are following the pioneers' lead and are bringing data-driven applications into the mainstream. Today, in addition to online channels, organizations' mobile and brick-and-mortar channels generate tremendous amounts of data, including machine, sensor, and other forms of IoT data. Retailers, telecommunications firms, companies in the entertainment industry, and others need big data analytics to drive externally facing data-driven applications that engage with customers and partners. Data-driven applications are also important for internal business management and process optimization.

Unlike traditional BI and analytics systems that separate analysis from processes, data-driven applications embed analytics to be in sync with application processes. Thus, data-driven applications need integrated systems that reduce the need for moving data to separate stores for transformation and analytics processing. Organizations should evaluate technologies that enable analytics to run natively where the data and database systems are, such as on Hadoop clusters or in cloud-based storage.

Data architectures must be flexible to support discovery, exploration, and experimentation with data; to gain competitive advantage, organizations need to innovate continuously in how applications apply analytical insights, particularly to respond in real time to customer behavior.

Data-driven applications must meet higher bars for concurrency, availability, resiliency, and security. They cannot function without their "brains" continuously providing direction through prescriptive analytics. Users will demand flexible options for visual data interaction so that they can examine the data and apply analytics to drill down to examine detailed data about specific customer segments or individual customers. In the case of cybersecurity, applications must enable security analysts to quickly identify cyber threats, effectively perform forensic analysis, and hunt for unknown actors across endpoints, networks, and users in one seamless system.

Organizations should develop a strategy for data-driven applications. They should prepare for big data analytics that are continuous, scalable, and highly available rather than running exclusively in off-hours batch processes. Organizations should plan for new use patterns, such as machine learning loops for improving application intelligence, and for frequent passes through the data to test analytics models.

Organizations should ensure that visual analytics enable users to work easily and flexibly to consume analytics, launch queries, and assemble personalized visualizations, dashboards, and applications.

NUMBER SIX

UNIFY ARCHITECTURE TO ENSURE PROPER GOVERNANCE, SECURITY, AND DATA COHERENCE

Establishing a unified architecture that incorporates diverse systems is never easy, but it is important. Organizations can gain greater business value from their data if they can align and integrate newer big data systems, cloud-based systems, and existing BI and data warehousing systems. With a unified architecture, organizations can more easily bring data together for users' analysis and reduce unnecessary duplication of both data and the processes being undertaken to prepare it.

When organizations have disparate systems and there is no unified architecture, individual groups tend to run all of their workloads on their respective systems, not thinking about whether the technology they have is really the best for that workload. Systems such as traditional data warehouses that were designed to collect and analyze historical data, for example, are not the best choice for analyzing streaming data.

A unified architecture can help organizations gain a big-picture view of their technologies and platforms so they can direct workloads to those that are fit for the purpose. In this way, a unified architecture enables organizations to improve overall performance, take pressure off systems that are not optimized for certain types of analytics workloads, and plan effectively for expansion.

Historically, big data systems based on open systems technologies have not been well integrated or unified. New technologies entered the environment without much planning for how they would fit together. In recent years, however, the Hadoop ecosystem has been evolving toward greater unity, with frameworks that support multiple processing and execution engines. By adhering to application programming interfaces (APIs) defined by the Apache Software Foundation, organizations can improve data flow from one application to another and enjoy greater portability. However, in most cases today, organizations still need to implement vendors' solutions to provide the level of unified management and integration they need.

Security and governance are two important reasons for unifying the architecture. Organizations should ensure that policies and rules are updated for big data analytics and are applied appropriately for each workload. Sensitive data must be secure and governed wherever it is stored and analyzed. If organizations are performing BI and analytics natively in the Hadoop ecosystem, they need management processes and technologies that enable them to secure and govern the data in place. Just as native big data analytics processing alleviates the need for data movement and development of intermediate systems, securing and governing data natively will also avoid unnecessary complexity and duplication of rules and policies inside and outside the Hadoop ecosystem.

Finally, metadata and semantic integration should be core objectives of a unified architecture. Users and automated data applications can then benefit from views of integrated data based on standard definitions, measures, metrics, and other semantics. It has traditionally been slow and labor-intensive to collect and standardize metadata and definitions, but tools are making the automation of steps easier. Organizations should evaluate technology capabilities for collection, cataloging, and integration of metadata.

FINAL WORD

This Checklist has discussed six focus areas for enabling easier, more effective access to big data for different types of analytics. Organizations should evaluate the emerging big data requirements from their users, who may be stymied by delays and difficulties with traditional BI and OLAP. New technologies have the potential to move users beyond those difficulties and facilitate projects that demand a broader range and larger volume of data.

Organizations should build a modern data platform strategy that encompasses the full range of use cases and technologies that can serve them. The modern vision must consider both on-premises and cloud-based big data; TDWI expects that cloud-based storage will take up an increasing percentage of available big data. Newer technologies can enable organizations to engage in smarter, faster, more visual access and analysis of big data sources; these sources are growing in importance to long-term strategic and daily business decisions.

ABOUT THE SPONSOR

ARCADIA DATA

arcadiadata.com

Arcadia Data provides the first data-native visual analytics platform to solve the most complex big data problems with the scale, agility, performance, and security users need to glean meaningful and real-time business insights in the era of big data. Arcadia Enterprise is a fully-distributed, massively parallel analytics platform purpose-built to allow business users, analysts, or data scientists to analyze large volumes of data down to the most granular level without moving data and build visual data applications scalable to hundreds and thousands of users. Arcadia Enterprise fills the gap between self-service BI and advanced analytics for use cases such as customer intelligence, connected car, cybersecurity, and trade surveillance. Arcadia Enterprise converges the visual, analytics, and data layers to provide accelerated access to all of the data stored within Hadoop, cloud, and other scale-out modern data platforms. The Arcadia Data platform is deployed by some of the world's leading brands, including Procter & Gamble, HPE, Royal Bank of Canada, Kaiser Permanente, and Neustar. To learn more, visit Arcadia Data at www.arcadiadata.com.

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.