10 BIG DATA CHALLENGES AND HOW TO ADDRESS THEM

Bringing a big data initiative to fruition requires an array of data skills and best practices. Here are 10 big data challenges enterprises must be ready for.

January 5, 2022 • TechTarget • George Lawton

A well-executed big data strategy can streamline operational costs, reduce time to market and enable new products. But enterprises face a variety of big data challenges in moving initiatives from boardroom discussions to practices that work.

IT and data professionals need to build out the physical infrastructure for moving data from different sources and between multiple applications. They also need to meet requirements for performance, scalability, timeliness, security and data governance. In addition, implementation costs must be considered upfront, as they can quickly spiral out of control.

Perhaps most importantly, enterprises need to figure out how and why big data matters to their business in the first place.

"One of the greatest challenges around big data projects comes down to successfully applying the insights captured," said Bill Szybillo, business intelligence manager at ERP software provider VAI.

Many applications and systems capture data, he explained, but organizations often struggle to understand what is valuable and, from there, to apply those insights in an impactful way.

Taking a broader look, here are 10 big data challenges that enterprises should be aware of and some pointers on how to address them.

1. Managing large volumes of data

Big data by its very definition typically involves large volumes of data housed in disparate systems and platforms. Szybillo said the first challenge for enterprises is consolidating the extremely large data sets they're extracting from CRM and ERP systems and other data sources into a unified and manageable big data architecture.

Once you have a sense of the data that's being collected, it becomes easier to narrow in on insights by making small adjustments, he said. To enable that, plan for an infrastructure that allows for incremental changes. Attempting big changes may just end up creating new problems.

2. Finding and fixing data quality issues

The analytics algorithms and artificial intelligence applications built on big data can generate bad results when data quality issues creep into big data systems. These problems can become more significant and harder to audit as data management and analytics teams attempt to pull in more and different types of data. Bunddler, an online marketplace for finding web shopping assistants who

help people buy products and arrange shipments, experienced these problems firsthand as it scaled to 500,000 customers. A key growth driver for the company was the use of big data to provide a highly personalized experience, reveal upselling opportunities and monitor new trends. Effective data quality management was a key concern.

"You need to monitor and fix any data quality issues constantly," Bunddler CEO Pavel Kovalenko said. Duplicate entries and typos are common, he said, especially when data comes from different sources. To ensure the quality of the data they collect, Kovalenko's team created an intelligent data identifier that matches duplicates with minor data variances and reports any possible typos. That has improved the accuracy of the business insights generated by analyzing the data.

3. Dealing with data integration and preparation complexities

Big data platforms solve the problem of collecting and storing large amounts of data of different types – and the quick retrieval of data that's needed for analytics uses. But the data collection process can still be very challenging, said Rosaria Silipo, a Ph.D. and principal data scientist at open source analytics platform vendor Knime.

The integrity of an enterprise's collected data stores is dependent on them being constantly updated. This requires maintaining access to a variety of data sources and having dedicated big data integration strategies.

Some enterprises use a data lake as a catch-all repository for sets of big data collected from diverse sources, without thinking through how the disparate data will be integrated. Various business domains, for example, produce data that is important for joint analysis, but this data often comes with different underlying semantics that must be disambiguated. Silipo cautions against ad hoc integration for projects, which can involve a lot of rework. For the optimal ROI on big data projects, it's generally better to develop a strategic approach to data integration.

4. Scaling big data systems efficiently and cost effectively

Enterprises can waste a lot of money storing big data if they don't have a strategy for how they want to use it. Organizations need to understand that big data analytics starts at the data ingestion stage, said George Kobakhidze, head of enterprise solutions at technology and services provider ZL Tech. Curating enterprise data repositories also requires consistent retention policies to cycle out old information, especially now because data that predates the COVID-19 pandemic is often no longer accurate in today's market.

Thus, data management teams should plan out the types, schemas and uses of data before deploying big data systems. But that's easier said than done, said Travis Rehl, vice president of product at cloud management platform vendor CloudCheckr.

"Oftentimes, you start from one data model and expand out but quickly realize the model doesn't fit your new data points and you suddenly have technical debt you need to resolve," he said.

A generic data lake with the appropriate data structure can make it easier to reuse data efficiently and cost effectively. For example, Parquet files often provide a better performance-to-cost ratio than CSV dumps within a data lake.

5. Evaluating and selecting big data technologies

Data management teams have a wide range of big data technologies to choose from, and the various tools often overlap in terms of their capabilities.

Lenley Hensarling, chief strategy officer at NoSQL database company Aerospike, recommends teams start by considering current and future needs for data from streaming and batch sources, such as mainframes, cloud applications and third-party data services. For example, enterprise-grade streaming platforms to consider include Apache Kafka, Apache Pulsar, AWS Kinesis and Google Pub/Sub – all of which provide seamless movement of data between cloud, on-premises and hybrid cloud systems, he said.

Next, teams should start evaluating the complex data preparation capabilities required to feed AI, machine learning and other advanced analytics systems. It's also important to plan for where the data might be processed. For circumstances where latency is an issue, teams need to consider how to run analytics and AI models on edge servers, and how to make it easy to update the models. These capabilities need to be balanced against the cost of deploying and managing the equipment and applications run on premises, in the cloud or on the edge.

6. Generating business insights

It's tempting for data teams to focus on the technology of big data, rather than outcomes. In many cases, Silipo has found that much less attention is placed on what to do with the data.

Generating valuable business insights from big data applications in organizations requires considering scenarios like creating KPI-based reports, identifying useful predictions or making different types of recommendations.

These efforts will require input from a mix of business analytics professionals, statisticians and data scientists with machine learning expertise. She said pairing that group with the big data engineering team can make a difference in increasing the ROI of setting up a big data environment.

7. Hiring and retaining workers with big data skills

"One of the biggest challenges regarding big data software development is finding and retaining the workers with big data skills," said Mike O'Malley, senior vice president of strategy at SenecaGlobal, a software development and IT outsourcing firm.

This particular big data trend isn't likely to go away soon. A report from S&P Global found that cloud architects and data scientists are among the most in-demand positions in 2021. One strategy for filling them is to partner with software development services companies that have already built out talent pools.

Another strategy is to work with HR to identify and address any gaps in existing big data talent, said Pablo Listingart, founder and owner of ComIT, a charity that provides free IT training.

"Many big data initiatives fail because of incorrect expectations and faulty estimations that are carried forward from the beginning of the project to the end," he said. The right team will be able to estimate risks, evaluate severity and resolve a variety of big data challenges.

It's also important to establish a culture for attracting and retaining the right talent. Vojtech Kurka, CTO at customer data platform vendor Meiro, said he started off imagining that he could solve every data problem with a few SQL and Python scripts in the right place. Over time, he realized he could get a lot further by hiring the right people and promoting a safe company culture that keeps people happy and motivated.

8. Keeping costs from getting out of control

Another common big data

challenge is what David Mariani, founder and CTO of data integration company AtScale, refers to as the "cloud bill heart attack." Many enterprises use existing data consumption metrics to estimate the costs of their new big data infrastructure – but that's a mistake.

One issue is that companies underestimate the sheer demand for computing resources that expanded access to richer data sets creates. The cloud in particular makes it easier for big data platforms to surface richer, more granular data, a capability that can drive up costs because cloud systems will elastically scale to meet user demand.

Using an on-demand pricing model can also increase costs. One good practice is to opt for fixed resource pricing, but that won't completely solve the problem. Although the meter stops at a fixed amount, poorly written applications may still end up eating resources that impact other users and workloads. So, another good practice lies in implementing fine-grained controls over queries. "I've seen several customers where users have written $10,000 queries due to poorly designed SQL," Mariani said.

CloudCheckr's Rehl also recommends that data management teams raise the cost issue upfront in their discussions with business and data engineering teams about big data deployments. It's the responsibility of the business to define what it is asking for; software developers should be responsible for delivering the data in an efficient format, and DevOps is responsible for ensuring the right archival policies and growth rates are monitored and managed.

9. Governing big data environments

Data governance issues become harder to address as big data applications grow across more systems. This problem is compounded as new cloud architectures enable enterprises to capture and store all the data they collect in its unaggregated form. Protected information fields can accidentally creep into a variety of applications.

"Without a data governance strategy and controls, much of the benefit of broader, deeper data access can be lost, in my experience," Mariani said.

A good practice is to treat data as a product, with built-in governance rules instituted from the beginning. Investing more time upfront in identifying and managing big data governance issues will make it easier to provide self-service access that doesn't require oversight of each new use case.

10. Ensuring data context and use cases are understood

Enterprises also tend to overemphasize the technology without understanding the context of the data and its uses for the business.

"There is often a ton of effort put into thinking about big data storage architectures, security frameworks and ingestion, but very little thought put into onboarding users and use cases," said Adam Wilson, CEO of data wrangling tools provider Trifacta.

Teams need to think about who will refine the data and how. Those closest to the business problems need to collaborate with those closest to the technology to manage risk and ensure proper alignment. This involves thinking about how to democratize the data engineering. It's also helpful to build out a few simple end-to-end use cases to get early wins, understand the limitations and engage users.