Solution White Paper

Why Automation Is Critical to Delivering Scalable and Reliable Big Data Solutions

Build, run, and manage complex data pipelines at scale with digital business automation

Table of Contents

Executive Summary

Executives have given their CIOs and business unit leaders a critical mission: Find ways to use insight-based analytics to support business transformation and create competitive advantages. Some version of that mission is in place at many enterprises, and the executive pressure behind it is strong. That leaves big data architects and their teams to struggle with deciding what data is needed, how it can be acquired, ingested, aggregated, processed, and managed so it can deliver the insights the business needs, and where existing data platforms fit in the changing environment. Big data isn't a project—It's a journey, and one that often comes without a roadmap.

Delivering big data capabilities with the scope and scale that enterprises need requires the flexibility to accommodate disparate data sources and work across disparate infrastructure, both on-premises and in the cloud. Doing it at the speed that executives and business conditions demand requires digital business automation. Control-M lets you automate essential tasks across the complete big data lifecycle, so you can coordinate and accelerate your digital business transformation. It is helping customers deliver new services to business users up to 20 percent faster while running their production environments with up to 25 percent fewer incidents and up to 50 percent less staff.

> **"As applications and services become more digitized, mobile, and interactive, the complexity of everyday transactions, workflows, and application architectures continues to escalate....Scalable and flexible workload automation solutions are critical to the success of these emerging applications."**

IDC, Worldwide Workload Management Software Market Shares, 2015: Year of Simplification and Self-Service, June 2016.

BIG DATA, BIG DECISIONS

Congratulations—you have executive-level management support to pursue big data projects, but you also have pressure to produce results quickly. You likely do not have direction on where to start, which technologies and tools to use, or how to architect the environment. Big data may be a project for those you report to, but for you, it's a journey. One of the great challenges is not to let all the details and decisions about architecture, tools, processes, and integration distract you from losing sight of how to deliver valuable insights and services to your business users.

CIOs, business unit leaders, enterprise architects, analysts, developers, and data scientists still need to work through a series of fundamental questions that must be answered before they can move the organization forward with big data:

- Is Hadoop the right framework for big data? And which distribution works best?
- Where do things like Pig, Hive, Spark, Mesos, NoSQL, Internet of Things (IoT) protocols, and other elements fit in?
- Where will we get the data we need?
- How can we get it into our systems?
- How can we operationalize machine learning and incorporate IoT?
- Can we maintain DevOps and continuous integration/continuous delivery (CI/CD) principles for our big data development?
- Where should the big data systems run: on-premises and/or in the cloud? What kind of cloud: private, public, or hybrid?

It is common for organizations to get bogged down with these questions and lose sight of the overall goal of creating systems that will provide better insight and improve decision making. The details are essential, but so is keeping focus on the big picture. The less planners need to focus on the details of how data will be ingested and how systems will run it, the more they can focus on finding new value and insight in their data.

PILOT PROJECTS: A DOUBLE-EDGED SWORD

Pilots give users the opportunity to provide a quick win by showing progress and delivering some results. However, pilots can also cover up problems that may surface later. For example, suppose a company wants to do a pilot to show how big data can create a 360-degree view of the customer. Much of the required data is already in the company's ERP and CRM systems (e.g., order history, total spending, points of contact, etc.), so data ingestion is not a problem that needs to be extensively studied at the pilot stage. Developers probably will not have all the tools they need to develop, test, and run the pilot service, but can do workarounds by writing scripts. **Management and development by scripting becomes unsustainable at high production volumes and speeds**. Data ingestion also changes at scale. In the example above, maintaining a 360-degree view of the entire customer base would need to extend data collection and analysis to social media, which would require the ability to efficiently handle streaming data. At a later stage, there may be a need to build models driven by machine learning that can allow the business to send customers promotions and recommendations based on preference and location.

The following sections present more insight about the challenges faced at different stages of the big data journey and how BMC can help.

PILOT TO PRODUCTION ISN'T A STEP. IT'S A LEAP.

When big data programs move from pilot to production, the data volumes get larger and the need to refresh data becomes more frequent. What may have been a one-off process for getting enterprise data into big data platforms like Hadoop now needs to repeat continually and flawlessly. New dependencies also emerge. Business users become more reliant on the outcome, so it becomes more important for reports and other jobs to run on-time.

For example, big data-driven recommendations may depend on analytics applied to daily transaction downloads. What happens if the file transfer fails to execute? The data itself may need to be filtered. Does the analytics workflow need all the input from an IoT sensor, or only interval samples, or only outlier results? Unstructured data many need to be formalized. These and other ETL-related activities can quickly consume all the time that data scientists have available, leaving them unable to do the valuable work of interpreting results and developing new use cases and business services.

The issues described above only relate to getting usable data into the system. The organization still has to create an infrastructure and workflows to process the data and deliver results. The most common problem that emerges at this point in the journey is when programs scale and the pilot environment does not match the production environment.

For example, the development tools used may not be compatible with the enterprise workflow management system, which would force the enterprise to maintain multiple job schedulers and scripts to manage different workflows. Big data pilot demonstrations may not have been developed according to enterprise security standards, thus would require more rounds of re-development and testing before they can be put into production. After those issues are resolved, there are dependencies between big data jobs and other processes, applications, and data sources that need to be synchronized and managed.

The differences between pilot and production environments represent real risks to big data success. Gartner predicts 60 percent of big data projects will fail to advance beyond the pilot stage.[1]

The processes for developing and deploying big data jobs need to be automated and repeatable. This becomes increasingly important as companies adopt DevOps to deliver new products and services faster. **High performing DevOps organizations are deploying an average of 200 times more applications than others.**[2] Workload organizations simply cannot afford to assign staff to do handholding and scripting every time a new service needs to be introduced. Otherwise, they will lose the rapid innovation advantage DevOps was intended to provide.

> **“Improving the relationship and processes between development and operations is certainly a good thing, but moving faster and changing more often can cause more breakage. Automation brings consistency and predictability.”**
>
> Dan Twing, EMA Radar™ for Workload Automation (WLA) Q1 2016: Report Summary and BMC Profile, March 2016.

AUTOMATE TO DRIVE SCALABILITY

Many of the aforementioned tasks can be automated with point solutions that come with the various components of big data technologies, such as the various open source tools that support the Hadoop environment (e.g., Oozie, Azkaban, Lipstick, Hue, Luigi, Pinball, and many more). Many of these tools perform as advertised but have limitations, and using too many can result in an environment marked by islands of automation with no end-to-end visibility. Tools also tend to be technology-specific, which can be limiting because the big data ecosystem is continually evolving.

For example, not long ago, most Hadoop environments used core elements like HDFS and MapReduce. Now Spark is commonly used as an alternative to MapReduce because it can run in memory and process jobs up to 1,000 times faster. Multiple NoSQL database options for Hadoop have been developed, including HBase, Cassandra, MongoDB, and Redis. There are also many options for bypassing or supplementing HDFS. If a strategic approach is not taken for automation, you could end up with a complicated mix of automation silos. Enterprises need to be careful to select the technologies and development approach that meet their current needs and preferences without locking them in to anything that will prevent them from taking advantage of future innovations.

> **“Digital business requires architectures that are purpose-built and flexible to adapt to an organization that expands its data and experiments with it.”**
>
> Rob van der Muelen, ”2017: The Year That Data and Analytics Go Mainstream.” Accessed 4/11/17 at http://www.gartner.com/smarterwithgartner/2017-the-year-that-data-and-analytics-go-mainstream/.

While automation is essential for running big data in production, automation systems must be able to support different technologies. BMC does that by providing flexible digital business automation for every stage of the big data journey.

[1] Gartner, ”Gartner Says Business Intelligence and Analytics Leaders Must Focus on Mindsets and Culture to Kick Start Advanced Analytics,” September 15, 2015, http://www.gartner.com/newsroom/id/3130017

[2] Puppet Labs and DORA (DevOps Research & Assessment), ”2016 State of DevOps Report,” https://puppet.com/resources/whitepaper/2016-state-of-devops-report

DATA INGESTION

One of the first steps is to bring together all the needed data into the big data system. Today, most processes are oriented to managing data from enterprise systems of record such as ERP and CRM solutions. Big data will require data from these systems and many other sources, including data warehouses, social media streams, machine-to-machine (M2M) interfaces, call center recordings, IoT solutions, and an expanding array of unstructured data. **Each of these data sources has its own methods and tools for managing and exchanging data, raising a risk that the big data infrastructure will actually be made up of a chain of islands of automation**. Hadoop open source tools don't completely solve the problem. For example, Sqoop is a leading open source data ingestion tool for Hadoop, but it only works within MapReduce and was developed for RDBMS data sources. Flume lets you ingest streaming data, but is limited for static sources. Other methods are available, but that creates more to learn and manage. These paths represent detours that delay progress on the big data journey.

Without a platform that can accommodate disparate data sources, enterprises will be forced to spend a lot of effort on integration, managing file transfers, and other tasks to overcome the lack of interoperability. Automation has a clear and valuable role in this scenario. When more big data and existing environments are consistent, more interactions between them can be automated.

CASE STUDY: NAVISTAR AND CONTROL-M

Navistar provides an excellent example of how automation improves the data ingestion process and creates real value for the enterprise. The manufacturer of commercial trucks, buses, defense vehicles, and engines launched OnCommand™, which is an IoT-based big data program that uses data collected from its vehicles to provide diagnostics services to fleet owners and suggest predictive maintenance. There are more than 250,000 trucks in Navistar's OnCommand program, and the company receives approximately 100 data points from each truck every day. To gain new insight for predictive maintenance, Navistar's Hadoop environment integrates input from sensors and other data sources from more than a dozen different telematics providers. **Customers have used the data and guidance to reduce their unplanned repairs and downtime by up to 30 percent**.

> **"In the past, the people working on big data projects were spending a lot of time manually moving data and running various kinds of scripts. It was time consuming and it sometimes meant they didn't have access to actionable data for days. They needed a faster, more efficient way to handle big data."**
>
> Todd Klessner, Senior Data Operations Specialist, Navistar
> ″Big Trucks Deliver Big Data at Navistar.″ Accessed 4/11/17 at http://www.bmc.com/blogs/big-trucks-deliver-big-data-navistar/.

Navistar previously had two engineers working full-time collecting data and formatting it for the organization′s Hadoop-based analytics program. Navistar was already using Control-M for workload automation and added Control-M for Hadoop to support its big data program.

Control-M provides native support for automated file transfers for big data and other environments. It can ingest file transfers from any source and integrate offloads and ingestion with other enterprise systems, including cloud-based sources. Control-M also supports Sqoop and the ETL functionality embedded in many leading big data and business intelligence solutions, including Cognos, Informatica, Oracle® Business Intelligence, SAP® Business Objects®, SQL Server SSIS, Cloudera, Hortonworks, MapR, and the IBM® BigInsights® distribution.

At Navistar, Control-M automated much of the data collection and security auditing tasks that used to be performed by engineers. They now spend their time developing new services instead of processing large amounts of incoming data. "We've just begun to tap the power of Control-M to help us use big data to enhance remote diagnostics, improve vehicle quality, and protect critical resources from unauthorized access, among other initiatives," said Todd Klessner, Senior Data Operations Specialist at Navistar.

DATA FORMATTING AND PROCESSING

Once data is ingested, the next steps in the journey are to get data aggregated and processed. The jobs and workflows that run here are what give context to the data in big data programs and turn it into something actionable. There are several challenges at this stage. First, enterprises have many development tools and approaches to choose from, but despite that, development is frequently a bottleneck. There are additional options when it comes to where big data processing will occur (e.g., HDFS, Hive, Spark, etc.), which influences decisions about the architecture and which tools to use.

Enterprises are not limited to one approach, but there is also the risk and complexity of managing multiple environments and creating islands of automation that are not interoperable. Development tools and test environments that are not seamlessly compatible with the production environment can seriously undermine the effectiveness of DevOps and other CI/CD approaches. For example, different point solutions may be convenient to use for developing associated workflows, but scheduling can be a time consuming job that requires a lot of scripting if the production team has to manage a lot of disparate, point-solution jobs. The task is especially challenging if jobs are dependent on each other. Such workflows are best handled through event-based triggering (e.g., the completion of one job automatically starts a dependent job), which is made harder if the jobs were developed with different tools.

Automation is extremely valuable at this stage because it can enable the short delivery cycles that DevOps is intended to produce. Rapid development is meaningless if new services only get delayed by promotion and scheduling after they are developed.

BMC's approach is to enable customers to develop and manage big data workflows the same way they do for other enterprise workflows. This approach creates consistency between the development, test, and production environments, so development and operations teams do not need to figure out who will handle various tasks for getting workflows into production. That minimizes the learning curve and the amount of proprietary toolsets the enterprise must support. Control-M does not just shift responsibilities for specific tasks between development and production. It automates the tasks so neither side has to do them, which saves a lot of time.

Big data workflows can be developed 40 percent faster with Control-M instead of using big data open source tools, an independent analysis found.[3]

BMC simplifies and automates big data development and execution in several ways:

- **Control-M enables big data jobs to be developed as code by embedding workflow automation in the application while it is being developed**. The Jobs-as-Code approach makes the development environment identical to the production environment and prevents many common failures and routine delays that occur when workflows are tested and promoted to production. Control-M creates an automated framework that makes it easier for development and operations teams to work together.
- **Control-M operationalizes machine learning.** Driving scalable, automated workflows is critical, as machine learning models are heavily dependent on the availability of data. However, it is common for data to have missing values and outliers. To make machine learning models effective, data must go through a series of preprocessing steps for finding, removing, and cleaning data from disparate sources in order to prepare it for machine learning. Control-M's ability to ingest and process data from any application and database makes it extremely effective in operationalizing machine learning from a single point of control.
- **Automation API lets users create big data workflows using their familiar, preferred development environment**. It is a set of programmatic interfaces (i.e., APIs and CLIs) that let developers and DevOps engineers use Control-M in a self-service manner within the agile application release process. Using JSON notation for job definitions, and GIT and RESTful APIs for validation, configuration, and deployment, workflow scheduling artifacts are seamlessly integrated with the enterprise's preferred tools used to automate the application release and deployment process.
- **Control-M automates multiple steps in the workflow development, testing, and promotion processes**. Examples include allowing workflows to be developed by drag-and-drop editing; automatically finding and changing naming conventions so workflows can be promoted from the test environment to production; automatically starting a workflow once a dependent job is complete (e.g., starting a job after a file transfer with data necessary for the job is completed); and enabling automated group updates.

[3] Impetus Technologies, "Apache Oozie vs. BMC Control-M Evaluation," June 7, 2016.

Automation for Big Data with Control-M

- Enable big data jobs to be developed as code
- Operationalize machine learning
- Create big data workflows in any environment
- Automate processes

DELIVERY

By now, the data has been ingested and automated workflows are coordinating the job scheduling, data transfers, and processing that turns your data into insight. What is needed next is to make sure that insight gets to the people and systems that can do something good with it. You should not have to make manual handoffs from the big data environment to data visualization and business intelligence applications (e.g., Cognos, Domo, Informatica, Tableau, Qlik, etc.). **Control-M helps by automating data movement, eliminating manual steps to move data into analytics systems, making sure SLAs are met by using predictive analytics to prevent job failures, and providing specific dashboards and self-service functions to different roles within your organization**.

As another customer example, GoPro provides mobile technology that helps people tell stories. The company's own experience helps tell the story of how Control-M helps organizations deliver their big data insights wherever they are needed. "GoPro's big data environment needs workload automation that natively integrates with Hadoop. More importantly, we need workload automation that spans inside and outside of Hadoop," said Joe Bentley, vice president of engineering at GoPro. "Control-M coordinates the automated processing of our big data. [The] commitment to natively integrate with Hadoop and other big data technologies is one of the drivers of our success."

Control-M doesn't just move data and workloads behind the scenes, but it also takes your big data across the finish line by delivering alerts, reports, and key insights directly to business users. Control-M can automate and manage output to business users through the web, mobile devices, print, and other outlets.

CONCLUSION

Control-M helps you automate every step of your big data project, including ingesting data to your systems, developing workflows to process it, and delivering results to business users and other systems that need to analyze the refined data. It also brings needed consistency and integration between big data and legacy environments. This ensures that big data will not be an island of automation, but rather a part of your digital business core. The benefit to this integration and automation is that you can innovate faster with less reliance on staff with specialized skills. Organizations that use Control-M can complete application delivery up to 20 percent faster and reduce production incidents by up to 25 percent.

FOR MORE INFORMATION

To learn more about how BMC supports big data and enables faster innovation, please visit **bmc.com/it-solutions/control-m-hadoop**