



Data Pipeline Automation

Create agile and efficient data pipelines with end-to-end automation

Executive Summary

Data is like oil. Its value is only realized when it's refined. This is a familiar maxim, but one that holds true for every organization that places data at the center of innovation and decision-making. Accurate, timely and complete data enables informed decisions, it determines customer experiences, sales process, governance strategy and much more.

However, many business leaders struggle to generate value from their data science projects. The reasons for this are many and varied. When the business asks for new reports, it can take weeks to create a new data source – by which time the insights may arrive too late. Moreover, many organizations devote a disproportionate amount of time **getting access** to the data, instead of **using** the data – tripped up by the familiar “keeping the lights on” versus “innovating” divide. Many also struggle with data preparation, sourcing or management. Without efficient access to data, there is no way to leverage machine learning (ML), artificial intelligence (AI) to support business innovation.

Automation is key to successful data management, alleviating data scientists and data engineers of the manual, repetitive data science tasks. But not all

automation strategies are equal. According to research by EMA, a majority of enterprises rely on a mix of automation tools, which diminishes control, increases the rate of failure and reduces agility.

The same study shows that 42% of organizations are having issues running application modernization strategies with their current automation solution. The complexity of integrating new technologies with legacy systems, automating hybrid processes from the mainframe to the cloud or leveraging more advanced analytics is simply a step too far for them.

The above explains why four-fifths of organizations are now looking to improve their automation strategies, in particular finding ways to get more centralized control over their automated processes.

This Broadcom eBook explores the organizational and automation challenges surrounding modern data pipelines. It also reveals how a data pipeline automation (DPA) strategy enables you to automate data pipelines end-to-end. By orchestrating tools, teams, infrastructure and data, DPA enables you to implement continuous innovation and analytics.

Data pipeline automation accelerates informed decision-making, innovation, customer success and business growth.

The challenges of modern data pipelines

Data is no longer the discarded, wasteful exhaust emanating out the back of operational systems. It is an essential ingredient to every business. As organizations pursue their digital transformation ambitions, data is at the heart of innovation, on everything from your people and customer strategy, to sales, finance and support.

Make no mistake, your ability to harness this data of ever-growing volume, variety and velocity will determine your future growth.

Here's the problem: it's not easy to manage that data pipeline efficiently. Too often, the data you need is generated in too many places and stored in too many silos. It's fragmented, disconnected and difficult to gather quickly. You've probably experienced the situation yourself. The business needs a new report drawn from new data sources. As a result, you need to integrate these additional data sources into your analytics pipeline. Many businesses struggle with these synchronization issues, which in turn make it harder to guarantee data consistency.

This helps explain why many IT organizations still devote up to 80% of their time getting the data rather than using the data. With artificial intelligence (AI) becoming a growing enterprise priority, data preparation and sourcing is expected to become a more significant pain point, impeding the ability to train machine learning (ML) algorithms to attain trust the models must provide.

To fully leverage the benefits of AI and ML, you have to adapt your analytics processes and data flows to move beyond the traditional data warehouse silos. The idea is to better integrate analytics developments with DevOps practices, where building, testing, provisioning and deployment are all run as automated processes. Emerging disciplines – such as DataOps – incorporate agile approaches to minimize the cycle time of analytics development. They aim to orchestrate environments, tools, models and data from an end-to-end perspective, bringing together data scientists with the operations teams to improve the data lifecycle management, and enable a fully efficient data pipeline.


However, to put DataOps to work, your organization first needs to address some specific organizational and automation challenges.



#1 Bridge technology and functional silos

Collaboration is essential. Teams, tools, infrastructure – they all need to be coordinated. Automation can bring all of these together, bridging technology and functional silos.

When you look at the vast number of new projects aimed to leverage the value of existing data, you quickly realize that many companies run big data environments in near-total isolation from the rest of enterprise business processes. Of course, there may be good reason for this: big data is a new technology that requires new skills that involve new teams. But running big data as a silo prevents you from integrating data-driven environments into your enterprise DevOps initiatives.



“Traditional workload automation strategies are unable to meet the needs of heterogeneous IT environments that include cloud-native infrastructure and big data workloads.”

— ***Gartner***¹

Big data projects are either too large or too complex to manage the traditional approach. This explains why early big data projects typically have poorly defined development processes. Waterfall approaches are notably inefficient as they deny access to a proper staging environment and enable limited time and scale for qualification. In other words, big data is implicitly promoting DevOps, because there is no real possibility to separate operations from development when you ultimately discover the relevance of your algorithms while in production.

The next generation of analytics will be the extension of DevOps and Continuous Delivery applied to the big data development. The notion is to provide ways to incorporate analytics developments into the agile model where building, testing, provisioning and deployment are all run as automated processes. The reality is that your data pipelines have to be automated alongside your continuous delivery pipelines. And for that, you need automation that goes beyond silos.

¹ Gartner, [Market Guide for Service Orchestration and Automation Platforms](#), Manjunath Bhat, Daniel Betts, Hassan Ennaciri, Chris Saunderson, April 17, 2020

#2 Remove manual handoffs and automate processes end-to-end

Your customers cannot wait for information. By automating data pipelines end-to-end, you reduce the potential for human errors, improve data quality and accelerate the data flows. Ultimately, you speed up the time to value from your data.

The amount of data that needs to be collected across an organization's business units, applications and external sources is growing exponentially. Additionally, more and more applications are hosted as a service in the cloud and integrated with big data instances such as Hadoop.

Point automation tools may be reliable and sufficiently scalable for low-volume, simple task scheduling. However, as workload volume or complexity increases, the exponential increase in manual effort to cope with this situation reveals its shortcomings.

Ensuring that information is available at the right time for strategic, tactical and operational purpose can be a challenge. Typical pipelines deal with extremely large volumes of data. Missing one step in the process, or executing a step at the wrong time, can result in a significant amount of wasted processing time – or, in the worst-case scenario, bad data.

The opportunistic ways of the past no longer deliver the requisite agility for a customer-responsive, application-centric business model. Manual handoffs and scripting between tools and teams create delays, bottlenecks and errors. A paradigm shift is required to meet business demands. You need automation that provides complete visibility and control across the entire data pipeline. By fully automating workflows that previously demanded manual intervention or synchronizations, you help build high-value tasks and process flows that run faster and more smoothly. This way, you free up valuable resources for more strategic projects, tightly binding your pipelines to your service level agreements (SLAs). It will not only improve IT agility and ensure consistent outcomes; it will also enable IT to speak the language of the business.

“60% of organizations have more than one WLA product in use.”

— Modernizing Workload Automation, Enterprise Management Associates® (EMA) 2020

#3 Enable data scientists and data engineers with self-service

Data scientists and data engineers remain heavily dependent on IT operations – to get access to the data, to provision compute environments, to deploy their work into production and for other tasks. Without automation, organizations risk creating a bottleneck, ultimately delaying the delivery of value from data.

On the one hand, big data offers a unique opportunity for faster, more informed decisions and personalized customer experiences. On the other, it brings the challenge of integrating new big data technologies without causing major disruption and impacting business-as-usual operations.

Data scientists find data flows too complex to design and manage, and subsequently call upon extended support from the IT experts. So, many IT organizations have difficulties in scaling with the volume of data, or the number of data sources, without slowing down development cycle times. As a result, innovation and customer experience are significantly impacted by the delays in getting access to data.

The main thrust of providing DPA must be to offer greater levels of accessibility to the right data, to as many people as possible. Data automation is needed to enable the complexities of big data to be hidden from individuals, while providing them with the data accessibility and insights they want.



“Democratize access to automation technologies to multiple groups within and outside IT by exposing self-service automation capabilities.”

— Gartner²

- ✓ Data scientists and data analysts that usually work in isolation from the big data engineers need to be on-boarded. This imposes higher levels of process standardization so that handoffs are seamless.
- ✓ The complexity of the underlying big data technology has to be hidden to the data scientists. This can be done by exposing clean, high-level APIs and predefined components that can be turned into simple graphical user interfaces to assemble workflows.
- ✓ Data scientists have to store their models, workflows and associated artifacts in the same repository that developers and other big data team members are using. This way, all application components can be easily orchestrated during deployment.
- ✓ Big data infrastructure has to be abstracted so that it can be pushed out in a continuous release fashion and deployed to any environment, whatever is it private, public, or hybrid cloud. The ideal state would resemble a 'Big-Data-as-a-Service' consumption model.

Consequences of not automating data pipelines

- ✓ **Customer experience:** Big data runs in isolation from enterprise processes
- ✓ **Speed:** Data scientists struggle to access data they need
- ✓ **Compliance:** Limited visibility and control on end-to-end data flows

#4 Integrate analytics into your enterprise business processes

Organizations need to capture, store and process data as it arrives, in any volume. And then distribute it to downstream applications, often in near real-time. Traditional data tools fall short in this regard, as you need to manage a mix of data movements and data processing. Automation meanwhile, improves the integration between data flows and traditional enterprise business processes.

As you work to do more and more with big data, it's only natural to expect your end-to-end business workflows to include an increasingly intricate blend of big data and traditional application jobs. Although this is certainly a normal result of incorporating big data into your broader workflows, it also means you'll have to contend with greater complexity as you work to simultaneously orchestrate big data and traditional jobs.

In most instances, big data users will have to run tasks separately from traditional ones. This greatly increases the time and effort required to deliver big data services to the business. Moreover, it limits your overarching visibility into all the operations – both traditional and big data – currently executing. And when this happens, it can create situations where confusion over the order in which tasks should be scheduled leads to slower response times and missed business opportunities.

Can open source tools automate data pipelines?

Open source tools can be used to automate data pipelines. Unlike holistic automation however, they only automate specific tasks on specific technologies. As a result, you create siloed islands of automation and risk losing control of the end-to-end pipeline.

“More than 40% of IT organizations struggle with AI because of data preparation, sourcing or management.”

— Forrester Infographic, AI Experiences A Reality Check”, May 17, 2019

Moreover, the deployment of analytics code and ML models need to be coordinated in parallel to the data pipeline, whichever open source or commercial tools are used. Here, automation can help your organization to improve the integration between modern data flows, and traditional enterprise business processes.

#5 Install high level of process standardization and governance

Regulations abound around data protection. Consequently, you need to set a strong governance, on complex, distributed processes. Automation can help you to create building blocks which you can reuse in your data flows. It means better standardization and better control on your processes.

Data pipelines reach across your company, your business partners, your supply chain, your SaaS offerings and more. It's a complex and diverse IT and business landscape. Visibility is typically limited to application silos or one-off integrations mandated by a partner, which means you are forced into blind handoffs of important data.

This lack of control over the entire enterprise process results in delays and errors that reveal themselves unexpectedly, and all too often when it's too late. But lack of visibility and control can also be the source of regulatory headaches. The EU General Data Protection Regulation (GDPR), for example, regulates all handling of personal data across industry segments and geo-trading zones.

Every country, state and industry has additional compliance measures that require detailed logs and audit trails of content encryption, access, authorization and usage. When you have a single, centralized point of control for data movements and policy management, you can quickly show your methodology and compliance to mitigate risks and avoid auditing headaches.

Key requirements for efficient data pipelines

- Bridge technology and functional silos
- Remove manual handoffs and automate processes end-to-end
- Enable data scientists and data engineers with self-service
- Integrate analytics with enterprise business processes
- Install high level of process standardization and governance

“32% of companies experienced audit and compliance concerns using non-WLA point automation tools.”

— *The Great Scheduler Migration*³

End-to-end data pipeline automation

Data science and analytics have become a critical component of enterprise decision-making. Now the business is calling out for timely, accurate and complete data to support innovation and growth. Your challenge is to accelerate and streamline that data delivery – whatever its volume, variety, or velocity – transforming source data into valuable insight.

The answer lies in data pipeline automation (DPA). This enterprise automation approach introduces repeatable processes – from the data teams, to IT, to the business – with automation that handles the scale and the volume of big data. By automating a mix of data movements and data processing in large volumes, DPA enables your organization to deliver valuable insight to the business front-line, more quickly. That results in more informed decision-making, enhanced customer experiences, agile sales and services strategies, and ultimately competitive advantage.

DPA automates data flows end-to-end. By orchestrating tools, teams, infrastructure and data, it enables your enterprise to deliver continuous innovation and analytics.



Figure 1 reveals how DPA automates all the steps to transform your vast volumes of source data into valuable insight.

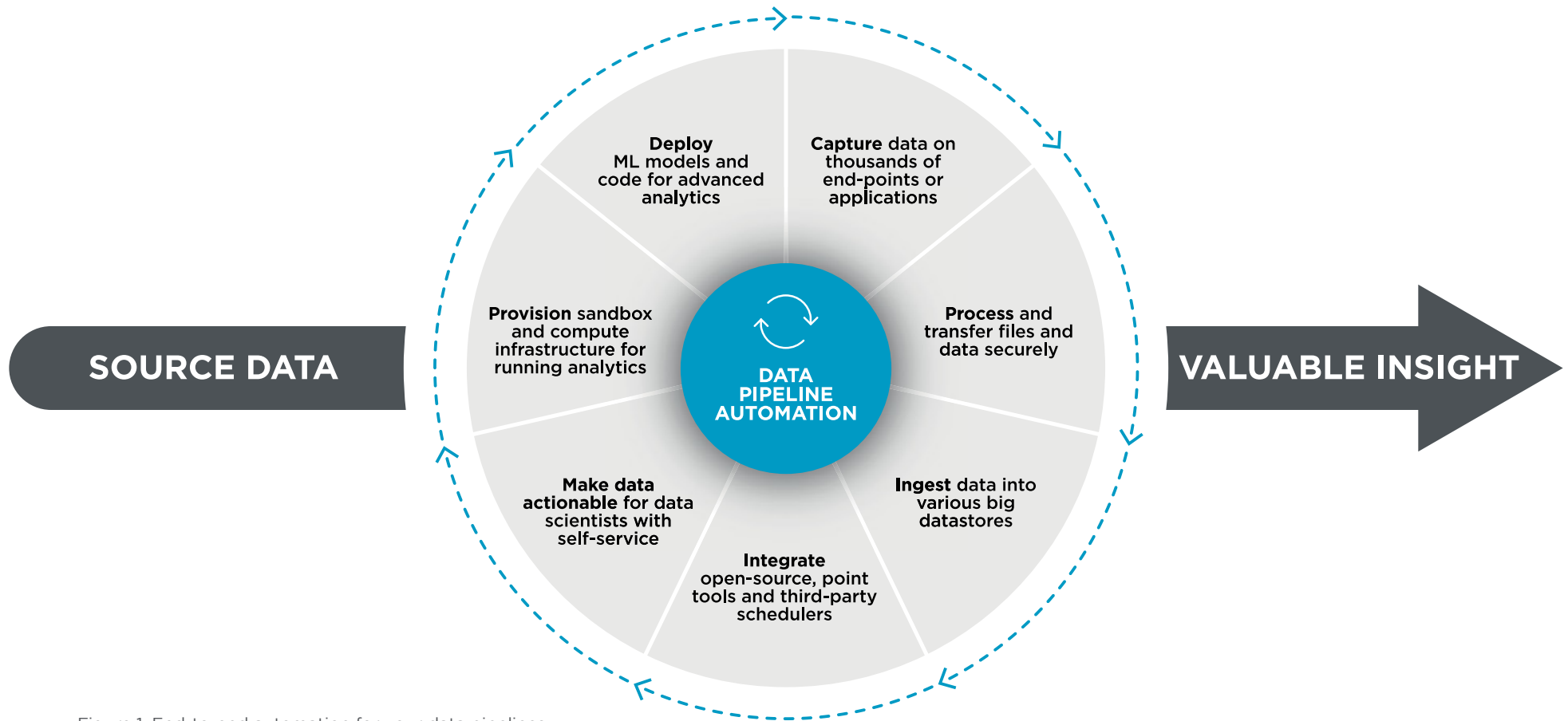


Figure 1: End-to-end automation for your data pipelines

- ✓ **Captures data** from thousands of end-points and applications
- ✓ **Processes** and transfers the files and data securely
- ✓ **Ingests** data into big data stores such as Hadoop
- ✓ **Integrates** with open source, point tools and third-party schedulers all along the data pipeline
- ✓ **Enables** self-service, to make data actionable for the data scientists
- ✓ **Provisions** sandbox and computing infrastructure for running analytics
- ✓ **Deploys** ML models and code for advanced analytics

DPA in action

A large U.S.-based utility company, providing electricity and gas to millions of households, is maximizing the benefits of smart metering with DPA. Typically, their smart meters record energy use as frequently as every half-hour and they report at least once daily – creating vast volumes of data. The utility provider's goal was to improve the 'meter to cash' process. The challenge was to integrate the corporate systems into this meter to cash process and to scale the process up to 86 million readings every day.

The solution was to automate the meter to cash process end-to-end. Using a Broadcom DPA solution, the organization automated that mix of data movements and data processing. As a result, the meter to cash process now runs 70% faster. And of course, with much greater reliability. Moreover, the seamless access to data has enabled innovation with the delivery of new digital services to customers. For example, real time monitoring of energy consumption, billing alerts, forecasting and other services are all now available on the customer portal.



Conclusion

Data is the fabric of your organization. It is the cornerstone of innovation, decision-making, customer experiences, sales and services strategy, compliance and much more. Your ability to manage efficient data pipelines has a direct impact on business success. By sharing accurate, timely and complete data analytics across the enterprise more quickly, you take a leap on your competitors and grow revenues.

Until recently however, the development of data and analytics pipelines, both simple and complex, has remained a handcrafted and largely non-repeatable process with minimal reuse, managed by data engineers working in isolation with different tools and approaches. The result is both a plodding development environment that can't keep pace with the demands of a data-driven business and an error-prone operational environment that is slow to respond to change requests.”

Data pipeline automation (DPA) changes all that. By orchestrating tools, teams, infrastructure and data through DPA, your organization is uniquely positioned to transform source data into valuable insight.



Broadcom Inc. is a global infrastructure technology leader built on 50 years of innovation, collaboration and engineering excellence.

Broadcom Inc. (NASDAQ: AVGO) is a global technology leader that designs, develops and supplies a broad range of semiconductor and infrastructure software solutions.

Broadcom's category-leading product portfolio serves critical markets including data center, networking, enterprise software, broadband, wireless, storage and industrial. Our solutions include data center networking and storage, enterprise and mainframe software focused on automation, monitoring and security, smartphone components, telecoms and factory automation. For more information, go to www.broadcom.com.

Learn more at:
broadcom.com/automation



For more information, visit our website at: www.broadcom.com

Copyright © 2023 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries. All trademarks, trade names, service marks, and logos referenced herein belong to their respective companies.
Data-Pipeline-Automation-eBook_BC-AUTO-2023_CE-3400_v5 April 7, 2023 8:26 am