

Why workflow orchestration is critical to delivering scalable and reliable data and analytics solutions

Build, run, and manage complex data pipelines at scale with workflow orchestration

Table of contents

- P3** Executive summary
- P4** Moving from data to decisions
- P5** Pilot projects: A double-edged sword
- P6** Pilot to production isn't a step. It's a leap
- P7** Orchestrate to drive scalability
- P10** Case study: Domino's Pizza and Control-M
- P11** Data formatting and processing
- P13** Delivery
- P14** Conclusion

Executive summary

CIOs and business unit leaders are on a critical mission to find ways to use insight-based analytics to support business transformation and create competitive advantage. Some version of that mission is already in place at many enterprises, and the executive pressure behind it is strong. As companies look to rapidly harness the power of artificial intelligence and machine learning (AI/ML), data architects and engineers have a vast and growing ecosystem of tools to choose from, adopt, and operationalize to deliver on those modern data initiatives. A good representation of this ecosystem is available in FirstMark's 2023 [MAD \(Machine Learning, Artificial Intelligence & Data\) Landscape](#) report. The industry response to this challenge has been to adopt new operational models such as DataOps and MLOps. In fact, the data community has started [The DataOps Manifesto](#) with guidance on several principles. One of those, which we will focus on here, is orchestration, a crucial building block in bringing together many moving parts of a data pipeline.

The [DataOps Manifesto](#) defines orchestrations as: "The beginning-to-end orchestration of data, tools, code, [and] environments." Control-M from BMC is the orchestration and operationalization platform that enables end-to-end orchestration of data pipelines and predictive service level agreements (SLAs) for any data technology or infrastructure across four key data stages: ingestion, storage, processing, and analytics. This whitepaper will delve into the orchestration and operationalization challenges faced by modern data teams and how Control-M addresses them.

Moving from data to decisions

Congratulations—you have executive-level management support to pursue data and analytics projects, but you also have pressure to produce results quickly. You are likely trying to figure out where to start, which technologies and tools to use, and how to architect the environment. Executives may see data as a project, 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 discovering 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 data:

- Do we have the right framework for to manage our data pipeline? And which distribution works best?
- Where do tools like Hadoop, Spark, EMR, Snowflake and RedShift fit in?
- Where will we get the data we need?
- How can we get it into our systems?
- How can we operationalize ML and

incorporate the Internet of Things (IoT)?

- Can we maintain DevOps and continuous integration and continuous delivery (CI/CD) principles for our data development?
- Where should the data systems run: on-premises, in the cloud, or both? 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 focusing 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 value and insight in their data.

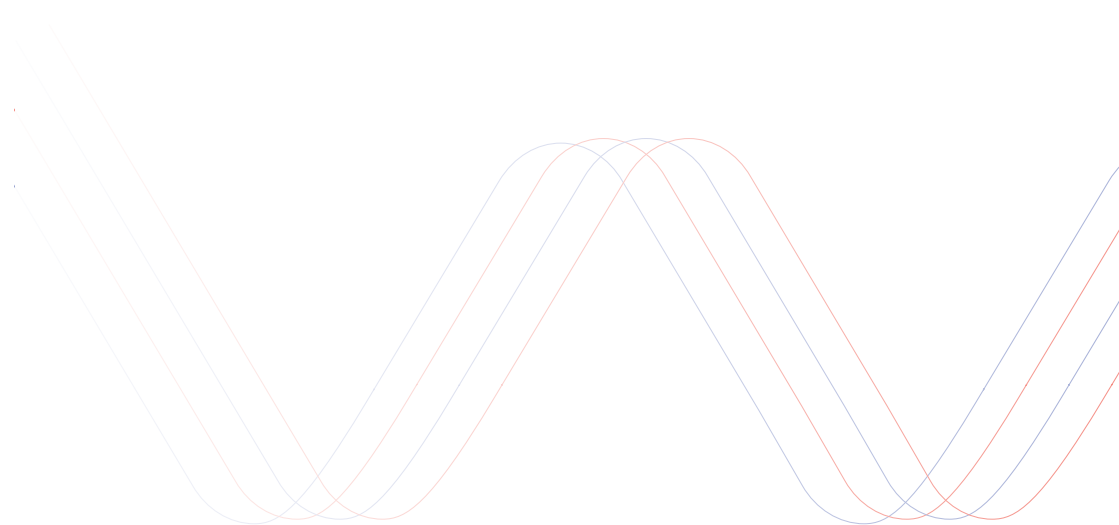
Pilot projects: A double-edged sword

Pilots give users the opportunity to provide a quick win by showing progress and delivering some results. However, pilots can also hide problems that may surface later. For example, suppose a company wants to do a pilot to show how data can create a 360-degree view of the customer. Much of the required data is already in the company's enterprise resource planning (ERP) and customer relationship management (CRM) systems make (i.e., 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 this instance, 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. And AI/ML models may be used 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 data and analytics journey and how Control-M can help.



Pilot to production isn't a step. It's a leap.

When data programs move from pilot to production, the 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 a data platform 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, 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 extract, transform, and load (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 must 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. Pilot demonstrations may not have been developed according to enterprise security standards, so they 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 data workflows 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 success.

The processes for developing and deploying data workflows need to be automated and repeatable. This becomes increasingly important as companies adopt DevOps to deliver new products and services faster. Workload teams 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.

Orchestrate to drive scalability

Many of the aforementioned tasks can be orchestrated with the point solutions included in data technologies, which perform as advertised but also have limitations. And using too many tools can create automation silos with no end-to-end visibility. Tools also tend to be technology-specific, which can be limiting because the data ecosystem is continually evolving.

If a strategic approach is not taken, you could end up with a complicated mix of automation silos. Enterprises need to be careful to select the technologies and development approaches that meet their current needs and preferences without locking them in to anything that will prevent them from taking advantage of future innovations.

While orchestration is essential for running data in production, automation systems must be able to support different technologies. Control-M does that by providing flexible workflow orchestration for every stage of the data and analytics journey.

Data ingestion

One of the first steps is to bring together all the needed data into your pipeline. Today, most processes are oriented to managing data from enterprise systems of record such as ERP and CRM solutions. Modern data workflows 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 data infrastructure will actually be comprised of a chain of islands of automation. Many application-specific tools don't completely solve the problem.

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. Orchestration has a clear and valuable role in this scenario. When more data and existing environments are consistent, more interactions between them can be orchestrated.

Essential workflow orchestration capabilities

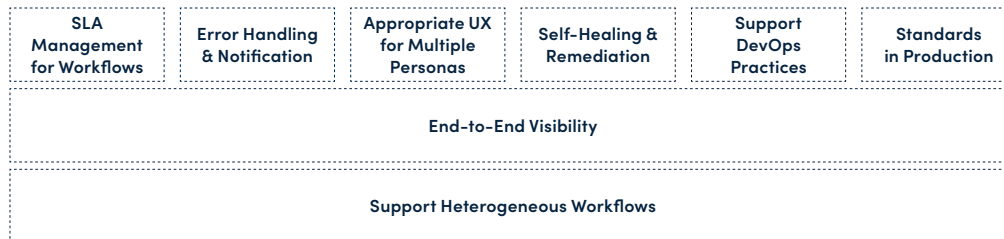


Figure 1. Eight essential workflow orchestration platform capabilities.

Successfully running business-critical workflows in production at scale doesn't happen by accident. The right workflow orchestration platform can help you streamline your data pipelines and get the actionable insights you need. That makes finding the right workflow orchestration platform vital.

With that in mind, here are eight essential capabilities to look for in your workflow orchestration platform:

Support heterogeneous workflows:

Companies are rapidly adopting cloud, and there is general agreement in the industry that the foreseeable future is one that is highly hybrid, a mix that can span from mainframe to distributed systems in the data center to multiple public and private clouds. If the orchestration platform cannot handle the diversity of applications and underlying infrastructure, then you will have a highly fragmented, siloed automation strategy that requires a lot of custom integration to handle cross-platform workflow dependencies.

SLA management: Business workflows ranging from ML models predicting risk to financial close and payment settlement can have completion SLAs that are governed by guidelines set by regulatory agencies. The orchestration platform must be able to understand and notify teams about failures and delays in the tasks that make up a complex workflow and link this to business impact.

Error handling and notification: When running in production, even the best-designed workflows will have failures and delays. It is vital that the right teams are notified so that lengthy war room discussions just to figure out who needs to work on a problem can be avoided. The orchestrator must enable notification to the right team at the right time.

Self-healing and remediation: When teams respond to job failures within business workflows, they take corrective action, such as restarting something, deleting a file, or flushing a cache or temp table. The orchestrator should allow automation

engineers to configure such actions to happen automatically the next time the same problem occurs, instead of just stopping a critical workflow in its tracks while several teams respond to the failures.

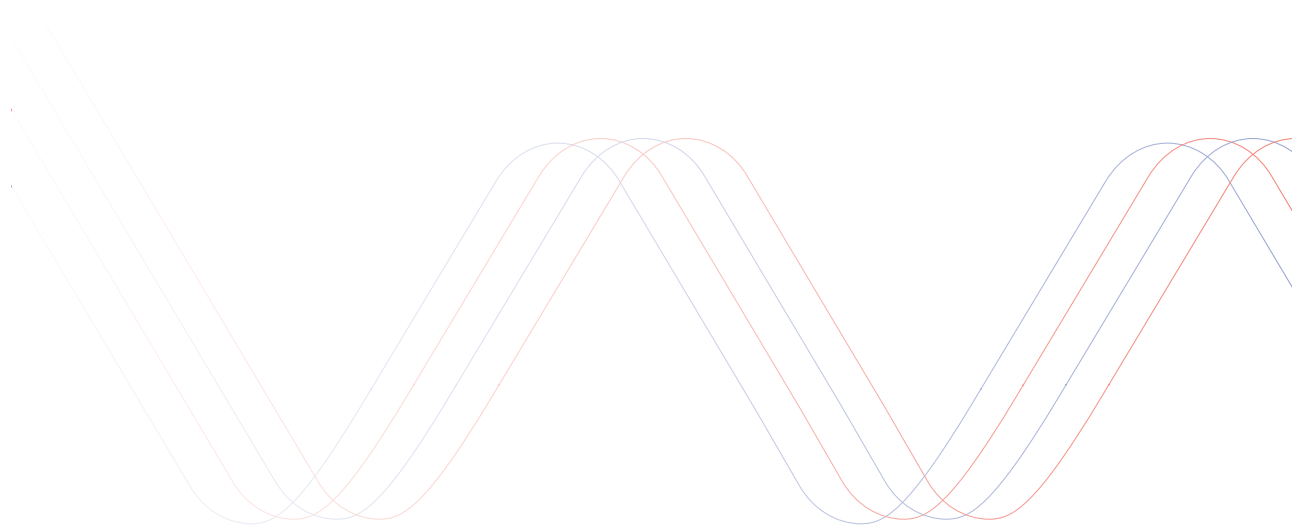
End-to-end visibility: While workflows execute interconnected business processes across a hybrid tech stack, the orchestration platform should be able to clearly show their lineage to better understand the relationships between applications and the business processes they support. This is also important for change management, to see what happens upstream and downstream from a process.

Appropriate UX for multiple personas:

Workflow orchestration is a team sport with many stakeholders such as developers, operations, and business process owners, etc. Each persona has a different use case in how they want to interact with the orchestrator. This means that the orchestrator must offer the right user interface and experience (UI/UX) for each persona so they can be effective users of the technology.

Standards in production: Running in production requires adherence to workflow standards such as correct naming conventions and error handling patterns, etc. The orchestration platform should have a very simple way to define them and guide users to the appropriate standards when they are building workflows.

Support DevOps practices: As companies adopt DevOps practices such as CI/CD pipelines, the development, modification, and even infrastructure deployment of the workflow orchestrator should fit into modern release practices.



Case study: Domino's Pizza and Control-M

[Domino's Pizza](#) provides an excellent example of how automation improves the data ingestion process and creates real value for the enterprise. Since 2007, it has focused on leveraging data and technology to create cutting-edge experiences that attract new customers and provide franchisees and executives with the analytics they need to make better business decisions faster.

Every day, Domino's funnels data from internal supply chain and staffing systems, sales, and orders from 16,000 U.S. and select international stores and third parties into more than 3,000 data pipelines and a data warehouse.

Getting data into the warehouse means ingesting, staging, and transforming it for use by multiple analytics pipelines, dashboards, reports, and other systems that drive decision-making on food quality and customer satisfaction, sharing important information back to franchisees so they can run their locations effectively.

Domino's needed a platform capable of orchestrating application and data workflows across an evolving tech stack—including many coordinated interactions with a complex mix of technologies, like MicroStrategy, AMQ, Apache Kafka, Confluent, GreenPlum,

Couchbase, TalendSQL Server, and Power BI. With rapidly growing data volumes and expanding business modernization initiatives, the platform had to be flexible enough to support long-term needs.

Control-M gives Domino's IT teams the robust operational capabilities and controls they need to ensure business services are delivered reliably and efficiently. The platform's growing list of integrations ensures they can bring in the technologies they need to continue to grow their business. Using Control-M, Domino's has orchestrated its expansive data pipelines, achieved significant growth in data volume, produced critical reports faster, improved SLA performance, and successfully rolled out new business services.

"Every single data attribute is critical to help us piece together the story and evaluate how we are doing. We always strive to outperform ourselves."

Deepti Soni

Senior IT Manager of Data Quality, Data Operations, and MLOps, Domino's Pizza

Data formatting and processing

Once data is ingested, the next steps in the journey are aggregation and processing. The workflows that run here are what give context to the data 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 data processing will occur (e.g., 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.,

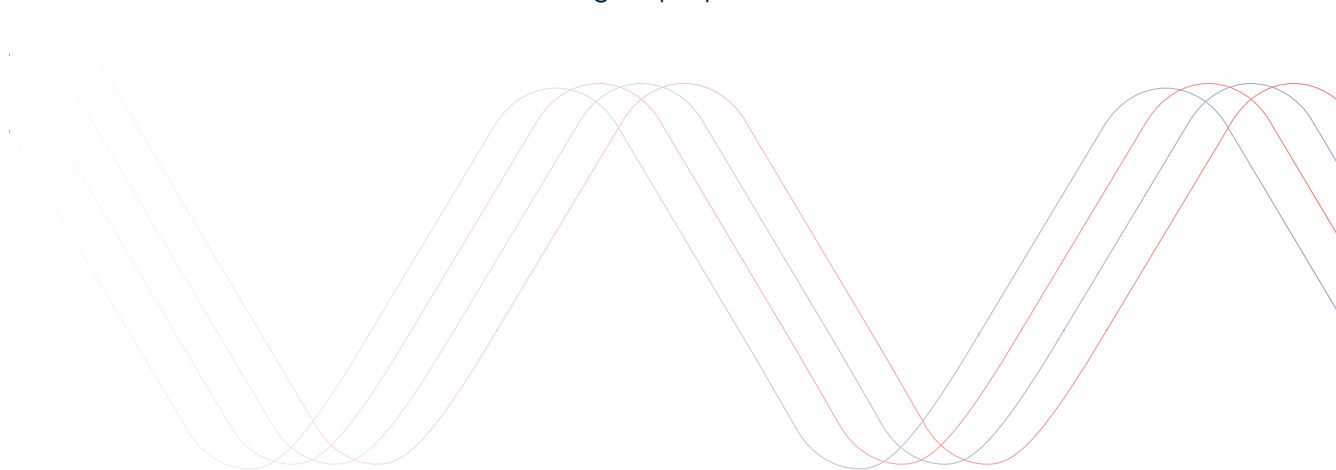
the completion of one job automatically starts a dependent job), which is made harder if the jobs were developed with different tools.

Orchestration 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.

Control-M enables customers to develop and manage 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 number of proprietary toolsets the enterprise must support. Control-M does not just shift responsibilities for specific tasks between development and production. It orchestrates the tasks so neither side has to do them, which saves a lot of time.

Control-M simplifies and orchestrates data development and execution in several ways:

- It enables data jobs to be developed as code by embedding workflow orchestration 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.
- It operationalizes ML. Driving scalable, automated workflows is critical, as ML models are heavily dependent on the availability of data. However, it is common for data to have missing values and outliers. To make ML models effective, data must go through a series of preprocessing steps for finding, removing, and cleaning information from disparate sources in order to prepare it for ML. Control-M's ability to ingest and process data from any application and database makes it extremely effective in operationalizing ML from a single point of control. It lets users create data workflows using their familiar, preferred development environment. Control-M Automation API 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.
- It orchestrates 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 they 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.



Delivery

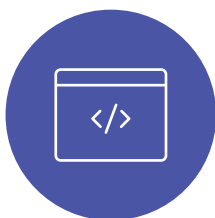
By now, the data has been ingested and orchestrated workflows are coordinating the job scheduling, data transfers, and processing that turns your data into insight. The next step is to make sure that insight gets to the people and systems that can do something good with it. You should not have to do manual handoffs from the data environment to visualization and business intelligence applications.

Control-M helps by orchestrating 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.

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

Data orchestration with Control-M



Enable data pipeline workflows to be developed as code



Operationalize AI/ML projects faster



Create data pipeline workflows in hybrid environments



Connect data workflows with enterprise application workflows

Conclusion

Control-M helps you orchestrate every step of your data and analytics 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 much-needed consistency to modern and legacy environments, allowing for better integration between them. This ensures that data will not be an island of automation, but rather a part of your digital core. The benefit to this integration and automation is that you can innovate faster with less reliance on staff with specialized skills.



For more information

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

About BMC

BMC empowers 86% of the Forbes Global 50 to transform digital operations into opportunity. Our leading portfolio of AI-enabled software connects data, automation, and observability across the business, enabling each customer to become an Autonomous Digital Enterprise ready to seize competitive advantage in a world of constant change.

www.bmc.com



BMC, the BMC logo, and BMC's other product names are the exclusive properties of BMC Software, Inc. or its affiliates, are registered or pending registration with the U.S. Patent and Trademark Office, and may be registered or pending registration in other countries. All other trademarks or registered trademarks are the property of their respective owners. ©Copyright 2024 BMC Software, Inc.

