Deliver Scalable Data and Analytics Solutions with Application Workflow Orchestration

Build, run, and manage complex data pipelines at scale

Table of Contents

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 advantages. Some version of that mission is in place at many enterprises, and the executive pressure behind it is strong. That leaves data architects and their teams to struggle with deciding what data is needed, how it can be acquired, ingested, aggregated, processed, and managed so they can deliver the insights the business needs. Data isn't a project—It's a journey, and one that often comes without a roadmap.

Delivering data and analytics capabilities with the scope and scale 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 executives and business conditions demand requires application workflow orchestration. This helps organizations orchestrate essential tasks across the complete data lifecycle, so they can coordinate and accelerate their digital transformation.

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. Those you report to 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 to manage our data pipelines? And which distribution works best?

- Where do tools like Hadoop, Spark, Amazon EMR, Snowflake and Amazon Redshift fit in?

- Where will we get the data we need?

- What are the best options for feeding new data streams into our systems? And, how can we leverage our existing systems of record?

- How can we operationalize machine learning and incorporate the Internet of Things (IoT)?

- Can we maintain DevOps and continuous integration/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 keeping focus on the big picture. The less planners need to focus on the details of how data will be ingested, 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 cover up 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 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 data and analytics journey and how BMC Helix Control-M, BMC's SaaS application workflow orchestration solution, can help.

Pilot to Production Stage Isn't a Step. It's a Leap.

When data programs move from pilot projects 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 may 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 the project needs to scale to production, but the tactics used in the pilot won't work in production.

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, thus would require more rounds of redevelopment 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, at scale. This becomes increasingly important when applying DevOps standards. **IT Ops 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 point solutions that come with the various components of data technologies. 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 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.

Orchestration is essential for running data in production and your chosen system must be able to support different technologies. BMC Helix 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. Data workflows often require information from systems of record (like ERPs and CRMs), 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 made up 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.

Case Study: Predictive Maintenance

A leading heavy equipment manufacturer that produces over-the-road trucks, engines and school buses for fleet owners and dealers worldwide launched an IoT-based program that uses data collected from its vehicles to provide diagnostics and suggested predictive maintenance services to fleet owners. There are more than 250,000 trucks that use the program, and the company inputs over 20 million data points into its platform every day. To gain new insight for predictive maintenance, the company 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 vehicle downtime by up to 40 percent.**

The company previously had two engineers working full-time collecting data and formatting it for the organization's Hadoop-based analytics program. They were already using application workflow orchestration in other parts of the business, and decided to use it to support the company's data and analytics program.

The company has been able to ingest file transfers from any source and integrate offloads and ingestion with other enterprise systems, including cloud-based sources. And they've been able to leverage 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, MapR, and the IBM® BigInsights® distribution.

The company has automated much of the data collection and security auditing tasks that used to be performed manually by engineers. They now spend their time developing new services instead of processing large amounts of incoming 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, and often because of 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.

BMC Helix 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. BMC Helix Control-M doesn't 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.

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

- It enables 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. BMC Helix Control-M creates an automated framework that makes it easier for development and operations teams to work together.

- It 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 information from disparate sources in order to prepare it for machine learning. BMC Helix 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.

- It lets users create data workflows using their familiar, preferred development environment. BMC Helix Control-M's Automation API is a set of programmatic interfaces (i.e., APIs and CLIs) that let developers and DevOps engineers use the solution 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. 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 data environment to visualization and business intelligence applications. **BMC Helix 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.**

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

Automation for Big Data with BMC Helix Control-M

Enable big data jobs to be developed as code

Operationalize machine learning

Create big data workflows in any environment

Automate processes

Conclusion

BMC Helix 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 needed consistency and integration between modern and legacy environments. 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 BMC Helix Control-M, please visit
www.bmc.com/helixcontrol-m

524945

WHITE PAPER

Deliver Scalable Data and Analytics Solutions with Application Workflow Orchestration

ABOUT [PARTNER]

Ut optatius et elit faccus molori veris dioris volore preriasim et es et labores expellupta cus et ut aut voluta temque poribus alit, quidel il iderrum rescipsus rende evendametur minciet hita et aut dolorit, simin perovitat rest dolorro rersperspe min cuptate ditate serae odi debis mollaccus enestrum latinve lignates est, cor suntor Ut optatius et elit faccus molori veris dioris volore preriasim et es et labores expellupta cus et ut aut voluta temque poribus alit, quidel il iderrum rescipsus rende evendametur minciet hita et aut dolorit, simin perovitat rest dolorro rersperspe min cuptate ditate serae odi debis mollaccus enestrum latinve lignates est, cor suntor