bmc

Solution White Paper

Connect Hadoop to the Enterprise

Streamline workflow automation with
BMC Control-M Application Integrator

Table of Contents

Executive Summary

The Apache® Hadoop® ecosystem has grown faster than the tools and knowledge available to harness it. Many enterprise big data programs have stalled as a result. Developers are struggling to turn the promise of big data into a practical reality. Dozens of available tools contribute to the challenge because they lack interoperability with existing enterprise systems and require specialized skills.

Organizations need a clear roadmap to negotiate the maze of proprietary tools, create big data workflows and put them into production. BMC Control-M Application Integrator, a workload automation design tool, helps organizations quickly deliver value from big data by creating new automated services.

INTRODUCTION

Big data may radically transform business, society, and life as we know it. Because interest is so high, there are almost as many opinions on how to reap these massive benefits as there are practitioners. As a result, the technology is highly dynamic.

Hadoop, the basic Apache project considered one of the fundamental components of big data, is made up of three modules: YARN, MapReduce, and HDFS. However, dozens of other Apache projects and commercial solutions are continually being developed and deployed.

As a stark example of this proliferation, the analyst firm Wikibon lists 46 vendors that get some or all of their revenue from software in the big data sector. Software solutions account for 22 percent of the entire market. In the open source world, 29 Apache projects are categorized as "big data" (see Apache's list here).

Many of these software offerings do much of their work in batch. The methods and tools for managing those batches are highly diverse. This situation poses a major challenge to developers building workflows that integrate this complex collection of tools, and to system administration/operations staff charged with keeping this hodgepodge running and fixing it when it breaks.

The challenges of managing a complex, heterogeneous collection of technology is not new or unique to big data. **However, the combination of relatively new and immature big data technology, the rate of change, the dearth of skills, and the pressing need to deliver business value faster make the problem especially onerous.**

The situation clearly requires a new approach for managing workflows.

THE UNDERLYING CONCEPT

If we deconstruct the process of managing workflow into its basic functions, there is broad common ground across the various options:

- **Access general scheduling and management capabilities that aren't specific to any particular application.** These include a user interface, a database to store information, auditing, a mechanism to perform functions at specific times or days, detecting events as triggers for action, managing events or actions in predecessor/successor relationships, and interacting with other management tools like email to notify service desks for incident management. All these capabilities are general and independent of specific applications. They are all necessary for a comprehensive solution, and if built correctly once, they should be reusable for all environments and applications that you wish to support.
- **Start a job.** This function can be unique to each application or environment. In Hadoop, a "job" consists of the Resource Manager passing containers to Application masters across nodes in the cluster; the entire collection is a single job. There is no way to start a job in Hadoop other than via Hadoop APIs (command line or web services), which submit a request to Yarn to start a job. Similar structures and concepts appear in many other applications where the unique job objects are internal to the application; there is a set of processes on an underlying OS or some other structure.

  The key for any application is some mechanism, such as Yarn, for requesting the initiation of something called a "job."
- **Monitor job progress and completion status.** This request provides information about the status of the job. Again, the application must initiate the request because there may not be any other component that understands what a "job" is within the particular application. In Hadoop, there is no way to determine the status of a job other than querying Yarn because information must be gathered from all the Application Managers running tasks throughout the cluster.
- **Terminate job execution.** Yet again, the function must be requested from Yarn.
- **Collect output.** In a simple application where a job is represented by a single process and all output is directed to standard output (stdout), collecting output may be a very simple process. However, in some applications, jobs write meaningful status to specific log files or even insert information into databases. To simplify problem analysis and retain meaningful job history, it would be desirable to extract this data from wherever the application stores it and place it into a standard location where all job output is managed by the workflow solution.

WHY NOT OOZIE?

If you are familiar with the Hadoop ecosystem, you may wonder why you wouldn't simply use Oozie or a similar tool to address this workflow challenge. After all, Oozie does for Hadoop/Yarn exactly what is described in the bullets above. So why not use Oozie? There is a three-part answer:

1. **Oozie only supports very specific Hadoop workflow actions** and a very limited set of applications within the broader Hadoop ecosystem (see Oozie actions here).
2. **Many workflow** (often referred to as "workload") c**apabilities have become de facto standards that Oozie does not support.** These include auditing, reporting, forecasting, output archiving, and many more.
3. **The Oozie interface seems intended for a very specific and highly technical audience** that is willing to invest substantial time manipulating XML documents and job property files.

Oozie clearly has limitations as a workflow management tool. An alternative providing proven, state-of-the-art workflow management functionality and the ability to be "taught" to perform all the required functions for a specific application in the Hadoop environment would provide substantial benefits.

WHY NOT USE NATIVE INTEGRATION FOR WORKLOAD AUTOMATION SOLUTIONS?

Outside the big data realm, enterprise workload automation solutions that provide native integration to different applications are the favored method for designing, scheduling, and monitoring traditional batch workflows. Consolidating workflow development and management enables people to develop new services more quickly and to manage workloads more effectively. Workload automation solutions also typically offer advanced alerting and troubleshooting features to support job execution, which are not available from application tools and would be time-consuming to develop through scripting. Using the graphical interface native to the enterprise scheduling solution is the best way to create and schedule jobs, and is preferred over scripting by nearly a 2:1 margin.[1]

A new hub for Hadoop help

The BMC Control-M Application Hub (www.bmc.com/hub) allows users to share the job types they've created with others and provides a forum for the Control-M community to exchange tips and ideas. The BMC Control-M Application Hub helps members quickly deliver new services by providing access to crowd-sourced integration job types. For example, the automated data exchange functions and complete backup workflow described above could all be posted to the Application Hub and made available to the community. The BMC Control-M Application Hub is an important resource for extending BMC Control-M workload automation.

BMC leads the workload automation product category with Control-M, a workload automation solution that automates batch services from a single point of control, accelerates delivery of digital services and increases quality of service. BMC Control-M provides a single tool set to develop, schedule, and manage mainframe, database, ERP, SQL, Java, web services, and many other job types. Users can create and schedule new business services with configuration options presented in the BMC Control-M interface, instead of having to develop code. The configure-don't-code approach saves time and prevents errors, resulting in faster deployments.

To further speed and simplify development, BMC provides integration to many popular software packages and platforms so they can be managed in the Control-M environment (visit www.bmc.com/it-solutions/control-m.html for a comprehensive listing). However, as the Hadoop ecosystem was exploding, native integration with Control-M was not available for the many flavors of Hadoop.

Native integration is the most convenient, cost-effective, and lowest-risk method for managing business services. The biggest limitation is the number of applications supported. Until now, customers have been dependent on the workload automation solution provider for integration to different applications. **With the introduction of BMC Control-M Application Integrator, Control-M users themselves can integrate any application with a command line.**

1 BMC 2015 AppDev Survey

BMC CONTROL-M APPLICATION INTEGRATOR IS THE NEXT GENERATION OF WORKLOAD AUTOMATION

BMC Control-M Application Integrator is a workload automation design tool that integrates the application process with BMC Control-M so business services are quickly and reliably delivered to customers. Backed by a community of users and crowd-sourced job types, organizations can quickly and easily deliver new or enhanced services in a scalable and sustainable way.

Unlike other workload automation tools requiring development of custom scripts that are costly to implement and maintain, **Application Integrator guides you through the steps for integrating any application with BMC Control-M to provide a single, enterprise view.**

With BMC Control-M Application Integrator, any application that has command line or web services APIs can be managed in the BMC Control-M environment. IT teams can create their own new business services, integrate and schedule them with other workflows, and monitor execution and SLAs, all through the familiar and convenient Control-M interface.

BMC Control-M Application Integrator saves development time by enabling users to create job forms that mirror the processing logic of their applications. BMC Control-M helps production control staff with quality assurance and improves on-time service delivery through features like simulation, advanced alerting, and automatic job restart functionality.

USING BMC CONTROL-M APPLICATION INTEGRATOR IN THE HADOOP ENVIRONMENT

BMC Control-M Application Integrator can, in effect, serve as a "scheduler of schedulers" for bringing big data solutions into the enterprise environment. This approach is analogous to the managers approach commonly used when a disparate collection of deployed tools must be controlled with a single, higher-level capability.

This section compares the way job types are created using Oozie and BMC Control-M Application Integrator.

OOZIE INTERFACES

Oozie provides web services and command-line interfaces. Command-line interfaces were chosen for this example because no programming is required and this entire tutorial can be quickly explored and tested by anyone with access to an operational Oozie environment.

Oozie usage information is displayed by running "Oozie help." A partial display of options for our tutorial is shown below.

Usage:

Oozie Help
Display usage for all commands or specified command

Oozie Version
Show Client Version

Oozie Job <options>
Job Operations

-config <arg>	Job configuration file '.xml' or '.properties'
-configcontent <arg>	Job configuration
-coordinator <arg>	Bundle rerun on coordinator names (requires -rerun)
-debug	Use debug mode to see debugging statements on stdout
-definition <arg>	Job definition
-doas <arg>	Do as user, impersonates as the specified user
-dryrun	Dryrun a workflow (since 3.3.2) or coordinator (since 2.0) job without actually executing it
-info <arg>	Information about a job
-kill <arg>	Kill a job
-oozie <arg>	Oozie URL
-rerun <arg>	Rerun a job (coordinator requires -action or -date, Bundle requires -coordinator or -date)
-run	Run a job

The very basic operations are:

- Run a job:

oozie job -run -oozie http://localhost:11000/oozie -config <path>/job.properties

This command return a job identifier following the "job: " string and appears similar to the following:
job: 0000002-150512021135238-oozie-user-W

- Query job status:

Use the job identifier to query the job status:
oozie job -oozie http://localhost:11000/oozie -info 0000002-150512021135238-oozie-user-W

- Kill a job:

Use the same job identifier to kill a job:
oozie job -oozie http://localhost:11000/oozie -kill 0000002-150512021135238-oozie-user-W

These three operations can perform the most basic functions of running a job, determining its success or failure, and terminating the job if necessary.

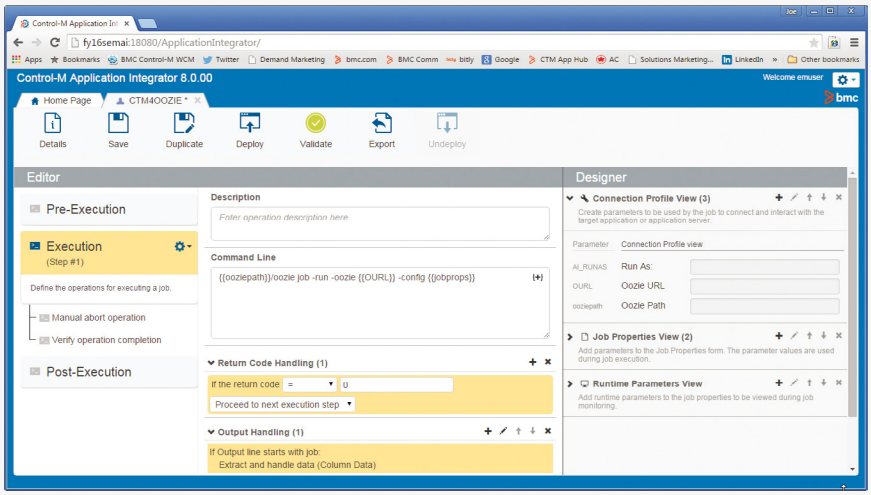
USING BMC CONTROL-M APPLICATION INTEGRATOR TO BUILD AND MANAGE JOBS

To provide a basis of comparison to Oozie, let's step through the process of creating and managing a job using BMC Control-M Application Integrator. Please keep in mind that although this example constructs an actual, executable job, the steps listed here are the bare minimum and are presented for illustration purposes only; in other words, don't try "just this" at home.

DESIGNING THE JOB TYPE

Using the BMC Control-M Application Integrator designer tool (Figure 1), we have created a Job Type named CTM4OOZIE.

FIGURE 1: The Control-M Application Integrator interface for creating jobs.



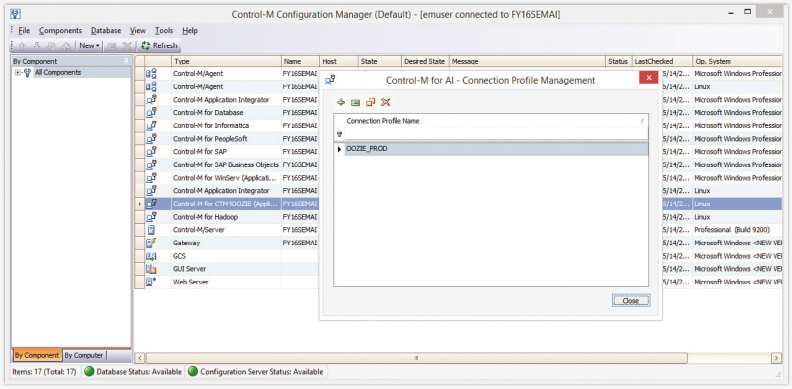
Notice the **Execution** branch in the "Editor" section on the left. That is what will execute when BMC Control-M is asked to run an Oozie job. It is followed by a "Manual abort operation" branch (which executes when a kill is requested) and a "Verify operation completed" branch, which will execute to check on the job status.

The "Designer" section on the right includes the Connection Profile view where application-level parameters can be defined.

After being created, the Job Type becomes available for use simply by using the "Deploy" function in the menu bar.

After the job deployment, BMC Control-M Configuration Manager is used to build a Connection Profile. This object, managed by Control-M, is where application-wide information used in multiple jobs is stored. Select the CTM4OOZIE application (which is the name of the job created in the previous step), and create the Connection Profile. One has already been defined in the screen shot in Figure 2.

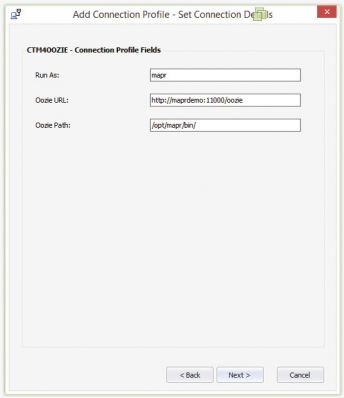
FIGURE 2: Creating Connection Profiles for various applications is simple.



The Connection Profile contains the values we designed. In our example, the values are:

- the user ID we will be running as or "impersonating"
- the URL endpoint for Oozie
- the path where the Oozie command line client is installed

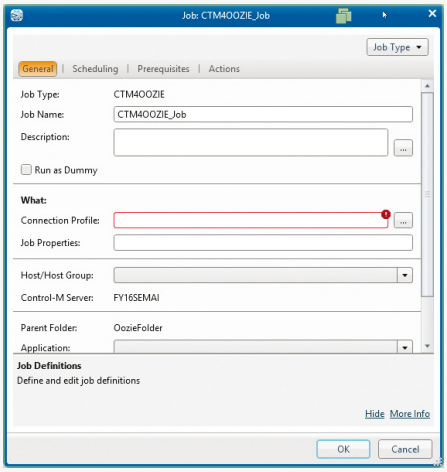
FIGURE 3: Each profile is given a name because multiple profiles can be defined. Here is how one appears:



BUILDING A JOB

Now let's examine how to build a job.

Once deployment was completed, the new job type was added to the job palette.



First, grab the desired job, drag it onto the canvas, and fill in the required information.

Next, supply the "config" parameter, which is usually a pointer to a job properties file, and the job is ready to run.

When BMC Control-M starts the job, the command defined in the Execution section of the Application Integrator editor is issued. It is similar to: "oozie job –run –oozie <URL from the connection profile> -config <properties from the job>"

Part of the Execution definition tells BMC Control-M to examine the command response for the job ID from a string similar to: "Job ID : 0000000-150514110252806-oozie-user-W" and passes that along.

To verify the job has completed, BMC Control-M executes the "verify operation completion" as defined, likely something similar to: "oozie job –run –oozie <URL from the connection profile> -info <Job ID>" from above.

Compared to Oozie and similar tools, BMC Control-M Application Integrator provides a simpler, more consistent approach for designing and deploying big data workflows. By streamlining the process, BMC **Control-M Application Integrator reduces the possibility of error when creating workflows, which helps organizations deliver new services to business users more quickly.** In turn, fast service delivery helps development organizations deliver on the promise of big data without driving up staff and support costs.

CONCLUSION

The top big data challenge for most organizations is to show value quickly. It is common for IT professionals to report spending 10 to 30 percent of a project's allocated time integrating different components. The time-consuming integration process can be frustrating and may spell the difference between success and failure.

Scripting and application-specific tools are not scalable or practical at a time when the focus is on speedy innovation. As new ecosystem components are built, an extensible platform that leverages existing skills and experience enables organizations to meet the challenge of working faster and delivering new services to the business.

BMC Control-M Application Integrator allows all BMC Control-M users to extend the reach and convenience of the automated workload scheduling solution to new environments, applications, and platforms. Following are the strategic benefits of BMC Control-M Application Integrator:

- For schedulers: provides an easy-to-use tool that operations or application development staff can use to quickly create new automated services.
- For IT leadership: enables faster introduction of new services and makes it more practical to support a large application portfolio.
- For executive leadership: leverages enterprise investments in Control-M and multiple applications, and enhances organizational agility by reducing the time needed to create new services.

BMC Control-M Application Integrator is a tool for the era of big data, which helps organizations improve service delivery and control costs.

FOR MORE INFORMATION

To learn more, please visit **www.bmc.com/integrate** and **www.bmc.com/hub.**

BMC is a global leader in software solutions that help IT transform traditional businesses into digital enterprises for the ultimate competitive advantage. Our Digital Enterprise Management set of IT solutions is designed to make digital business fast, seamless, and optimized. From mainframe to mobile to cloud and beyond, we pair high-speed digital innovation with robust IT industrialization—allowing our customers to provide intuitive user experiences with optimized performance, cost, compliance, and productivity. BMC solutions serve more than 15,000 customers worldwide including 82 percent of the Fortune 500®.

BMC – Bring IT to Life

bmc

BMC, BMC Software, the BMC logo, and the BMC Software logo, and all other BMC Software product and service names are owned by BMC Software, Inc. and are registered or pending registration in the US Patent and Trademark Office or in the trademark offices of other countries. All other trademarks belong to their respective companies. © Copyright 2015 BMC Software, Inc.

469219