bmc | IMPETUS

Solution White Paper

Comparing Hadoop Open Source Tools and Control-M for Workflow Management

Discover the digital enterprise automation capabilities of Control-M versus Apache Oozie, and Hue

Table of Contents

Executive Summary

Hadoop® has become the preferred platform for enterprise big data programs, but many enterprises are still wrestling with fundamental decisions about how they should implement it. One of the biggest and most persistent questions is whether to use Hadoop open source tools or established enterprise resources for developing and managing Hadoop services.

The goal of this white paper is to compare Hadoop open source tools, including Hue and Oozie, with Control-M from BMC for the common tasks of creating, testing, deploying, and managing a Hadoop-based workflow. Impetus Technologies, a big data software products and services company with extensive experience in developing and implementing Hadoop solutions for Fortune 1000 companies, was commissioned to develop and deploy a user monitoring mobile application using each tool.

The results of the testing showed that it took approximately **40 percent less time to complete workflow development using Control-M than with the traditional method of using Hadoop open source tools.** This white paper describes the tests and explains what specific features and functionalities that Impetus found valuable for the Hadoop environment.

Key Findings

- Development of workflows with Control-M was 40% faster than Oozie
- Running workflows via Control-M is more secure as it does not expose login credentials for tasks such as database extracts and file transfers
- Control-M allows managing Hadoop as well as non-Hadoop workflows from a single interface allowing scalability and simplifying operations

Table 1: Key Differences Found Between Hadoop Open Source Tools and Control-M for Hadoop®

Functionality	Hadoop Open Source Tools	Control-M for Hadoop
View Hadoop and non-Hadoop jobs from one interface	No	Yes
Change workload schedules because of holidays, exceptions	Done by scripting	Yes
Tag jobs with priority levels for dynamic execution	No	Yes
Automatically promote workflows between environments	No	Yes
Create dependencies between workflows	No	Yes
Native support for FTP and SFTP	No	Yes
Data encryption	No	PGP, SSL
Automatically restart failed job transfers	No	Yes
Interfaces to ticketing systems	No	Yes
Report job status to mobile devices	No	Yes

ABOUT THE PROJECT

Impetus was asked to lead the project because of its extensive experience. Impetus has more than 800 clients and has completed numerous Hadoop projects involving technology strategy, solution architecture, proof of concept, production implementation, and ongoing support.

Impetus evaluated Hadoop open source tools and Control-M in five areas that are common to nearly all big data projects:

1. Building workflows
2. Scheduling and managing jobs
3. Conducting data imports and file transfers
4. Updating jobs
5. Security

Impetus chose a use case that is common to many of its clients and other organizations that are using big data. It created a workflow to identify users of an enterprise mobile application and track their interactions with the app.

Three aspects of the participation by Impetus are notable:

- Oozie is the standard workload scheduling tool for Hadoop projects. Impetus staff has worked with Oozie in dozens of previous engagements.
- Impetus staff had no previous experience with Control-M prior to the test project.
- BMC provided minimal technical support to Impetus. The documentation that comes with Control-M was their main means of learning how to use the solution. The companies held weekly project update calls. BMC answered questions during the calls, but did not provide any formal training.

Use Case – Capture Data on Mobile App Users

The mobile app analysis use case required five general steps:

1. Build Hadoop workflows to create the use case
2. Promote workflows from development to test, and from test to production environments
3. Schedule the workflows to run in the production environment, including event-based scheduling
4. Integrate file transfers to provide the data on which the workflows operate and export results
5. Complete a mass update of all jobs

The mobile application's daily logs were processed to capture the ZIP code for each user, all interactions by each user, and all interactions at the ZIP code level (Figure 1 presents the flow chart for the use case). The input data for the workflow consisted of:

- A log file with time stamp, user ID, and state, formatted as a text file separated by space
- Lookup data consisting of user ID, age, and gender, in a database table format
- Lookup data consisting of state and ZIP code in a database table

The workflow required processed data to be output in three ways:

1. Hive tables
2. RDBMS
3. Streaming via Kafka Topic

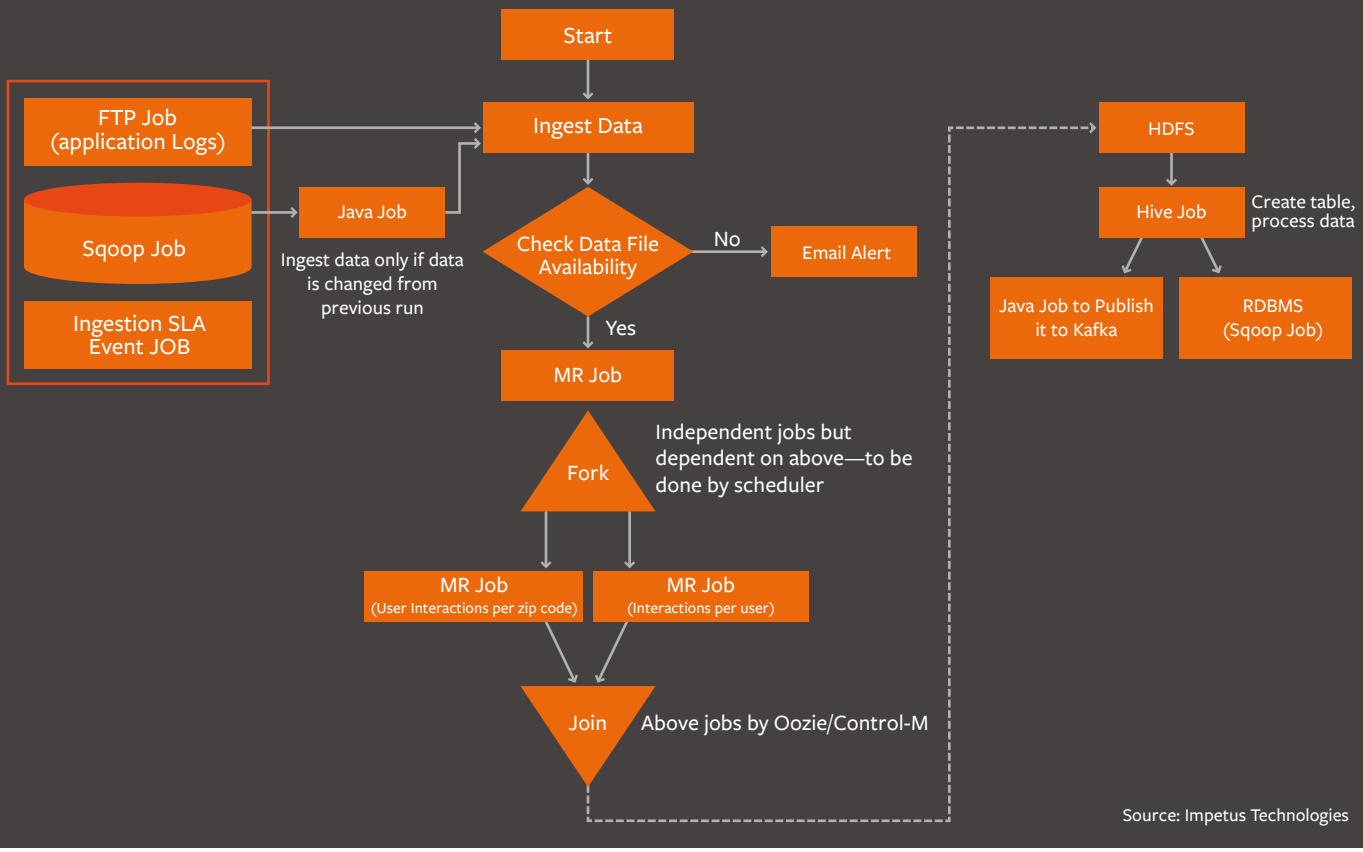

FIGURE 1: Mobile Application Interaction Tracking Use Case Workflow

System Configuration

A virtualized Linux® environment running the Cloudera Hadoop distribution was used for workload development and testing. Popular, common elements of the Hadoop ecosystem were used so the project mirrored many of the real-world development tools and techniques that enterprises use, including Oozie, Hue, Pig, Hive, Sqoop, and Impala.

The following sections describe the tasks Impetus needed to perform in order to complete the Hadoop open source and Control-M mobile application workflows; key differences between the methods; and other observations about each solution's suitability for different Hadoop development and execution scenarios.

BUILDING WORKFLOWS

Building workflows is an essential activity and also represents one of the most fundamental differences between Control-M and Hadoop tools. The comparison considered things like how workflows are actually developed and debugged, how convenient and intuitive each method is, the method for importing and exporting jobs, and how version control is managed.

For the open source build with Oozie, Impetus engineers created workflows using the Hue graphical interface, wrote Java® programs to incorporate Kafka® and various data mappers and reducers, and used a combination of Hue web interfaces or XML interfaces they wrote for the job. After the workflows were designed and tested, Oozie scheduled those workflows.

Here are some of the Impetus developers' notes and comments[1] about the workflow building process:

Oozie workflows definition can be provided in a XML file or can be created using Oozie CLI or Hue UI. All constructs supported by Oozie can be composed via the XML file approach, however neither the Oozie UI or Hue seem to be mature enough to match the manual/XML approach.

- The Oozie UI does not support creation of jobs
- Hue UI (3.7) does not support scheduling of jobs based on data-events; import/export button is missing
- Hue UI (3.7) workflow crashed when transitioning is done for decision jobs

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

Control-M Workflow Development

One of the main differences in the Control-M environment is that it allows, but does not require, development by command line interfaces (CLI). Control-M gives users the choice of developing using command line interfaces or using its native GUI that supports drag-and-drop development and includes wizards and drop-down menus for common tasks. Both GUI and CLI development can be used in the same workflow; however Impetus worked exclusively through GUI.

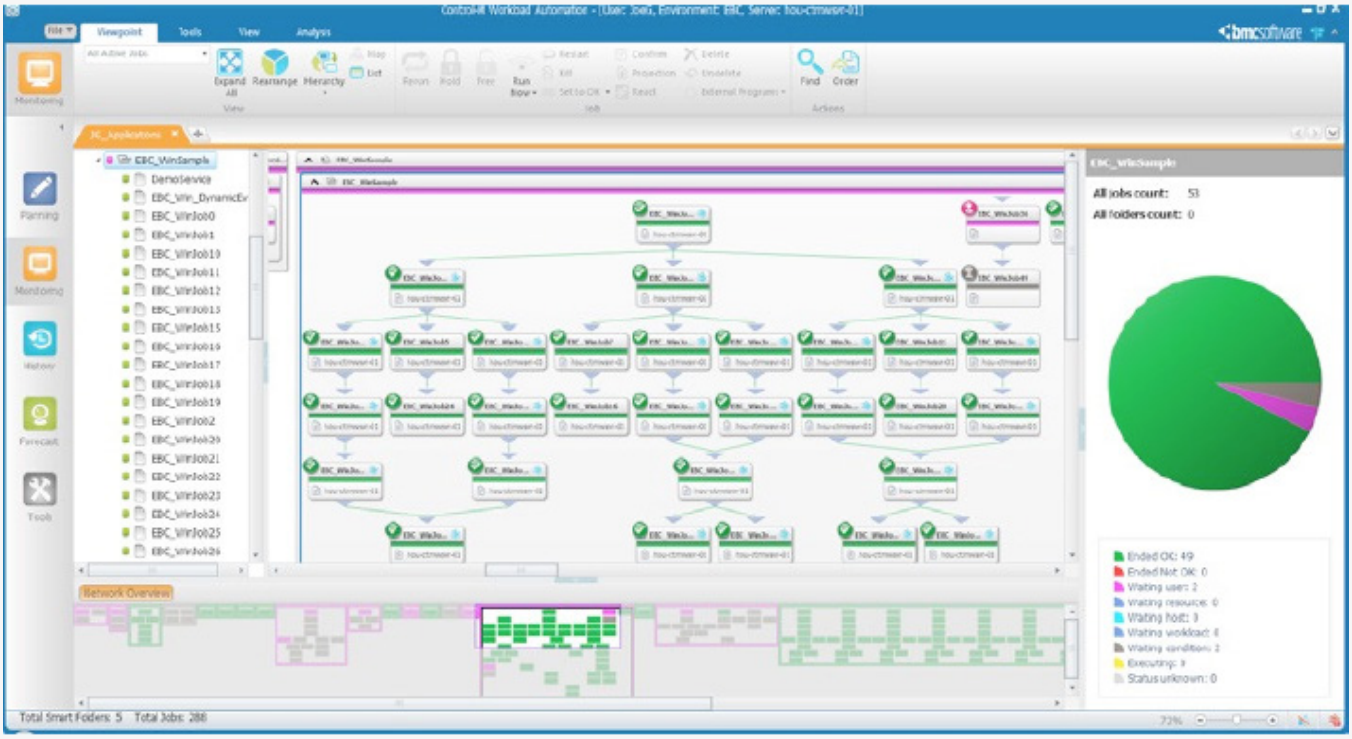

FIGURE 2: The Control-M GUI for Hadoop development

The Impetus engineers especially valued the functionality of Control-M that lets users tag jobs and define critical paths. As noted, the Impetus staff had no prior experience with Control-M and received no training from BMC. However, Impetus reported that the Control-M GUI simplified development. **It took approximately 40 percent less time to complete workflow development using Control-M than with Oozie and other Hadoop open source tools**. Control-M also accomplished several tasks that Oozie is incapable of such as handling encryption out-of-the-box for file transfers, SLA management, and critical path management for Hadoop and non-Hadoop workflows. Attempting to incorporate this in Oozie would have required a significant amount of scripting, which would have added weeks to development and testing.

Much of the savings in time for development was attributed to Control-M consolidating and automating many specific tasks that were fundamental to workflow development. Impetus noted how this consolidation and automation saved time, which supported a finding from a separate Hadoop workflow automation study by Forrester Consulting:

"Although you can run Hadoop without workflows without an enterprise grade automation solution, it'll take you more time and effort to streamline your deployment. Without automation, most enterprises spend months of effort in setting up, installing, configuring, and managing Hadoop clusters."[2]

SCHEDULING WORKFLOWS

Oozie is a Hadoop-specific scheduling utility and only supports specific Hadoop functions. Because Oozie can only be used for Hadoop jobs, other workload schedulers are commonly used instead of or in addition to Oozie in Hadoop environments. Control-M can be used to manage Hadoop and non-Hadoop workloads. The Impetus scheduling comparison focused only on Hadoop workloads. It found three significant differences.

2 Forrester Consulting, "Hadoop Workflow Automation Improves Efficiency and Increases Competitive Advantage," October 2015.

Feature	Control-M	Oozie	Explanation
Dependency management	✓		Oozie uses coordinators, which are collections of jobs. It does not provide a way to create a dependency between one coordinator and another. Dependencies can be created in Control-M simply by clicking on the workflows.
Enhanced calendar functionality	✓		Exception handling, such as having a job run at the same time every day except when the day is a holiday, is also done differently in Oozie and Control-M. Oozie primarily handles exception scheduling by having developers create XML or scripts. Control-M users can make the change through a calendar-based schedule.
Time zone conversion	✓		Oozie defaults to UTC, so users that don't keep time in UTC need to convert jobs to run at the desired time for the specific time zone. Control-M users can set and change the time zone for job execution from a pull-down menu.
Scheduling jobs to run at specific times of day	✓	✓	Both meet this requirement.
Scheduling jobs to run at regular intervals	✓	✓	Both meet this requirement.
Contingent scheduling based on predecessor/ successor relationships	✓	✓	Both meet this requirement. (ex: scheduling Job B to run immediately after Job A is finished)
File watchers that monitor for a file event such as file creation or deletion	✓	✓	Both meet this requirement, but Oozie's file watcher is limited to the Hadoop Distributed File System (HDFS) whereas Control-M can watch for files on HDFS and conventional file systems, including remote file systems that are only accessible via FTP and SSH servers.

EXECUTING AND MANAGING WORKFLOWS

A workflow is typically built in a test or development environment and a separate action is required to move it into the production environment where it can execute. The job then needs to be monitored to make sure it runs correctly, especially if other workflows depend on it. Any problems need to be identified and corrected so jobs can continue. Workload scheduling plays an important role in all these activities, so Impetus compared how Oozie and Control-M each contribute to supporting reliable workflow execution.

Initiating Jobs

Many Hadoop workflows are not made up exclusively of Hadoop jobs—they depend on other applications and contain jobs that execute "traditional" applications that require Hadoop systems to interact with the outside world. Impetus tested how each approach would manage a situation where a Hadoop job started data ingestion from a file transfer from another system. Control-M displays Hadoop and non-Hadoop workflows and their dependencies in a single interface. Oozie only displays Hadoop workflows, so users needed to check multiple systems to determine job status.

Promoting Jobs and Other Job Movements Between Environments

Workflows cannot be created and put into production without going through several other processes first. For example, they are usually tested in multiple environments before being put into production. Promoting workflows to production and making other moves between environments (e.g., development to test, test to QA, sending jobs back to development for additional work, etc.) are not simple tasks in the core Hadoop environment. Impetus identified at least two Control-M features that make promotion/ demotion easier and have no equivalent in Oozie:

- The attribute changes that must occur to promote workflows (e.g., changing naming conventions) must be done manually in the Hadoop environment, which does not allow users to run a find/replace function to make repetitive changes. The Control-M GUI automates name changes and many other repetitive tasks that are routine to promoting jobs.
- For additional quality control, Control-M can be set to automatically find jobs that need to be updated and present them for action.

Managing File Transfers

File transfers are a common way to move data in and out of Hadoop environments. Therefore, the file transfer functionality and support that organizations build into their Hadoop environment are important variables to how reliable, successful, and easy to manage the environment will be.

FTP and SFTP Support

- Most notably, Oozie does not provide direct support for FTP and SFTP. To use those protocols in Oozie, development in Java or scripting is needed, plus additional development if encryption is desired.
- Control-M natively supports FTP and SFTP and encryption, and provides a built-in file transfer server, plus options for FTP over SSL and PGP encryption.

Recovering Failed Jobs

- If job transfers fail, Oozie does not automatically adjust schedules or make any provisions to facilitate recovery and preserve business continuity.
- Control-M automatically attempts to restart incomplete transfers from the point of failure. Control-M also allows users to build in business logic in case a file arrives late, so scheduling can be adjusted to minimize bottlenecks.

Managing Workflows

Workflow scheduling is not a "set it and forget it" task. Schedules regularly need to be changed based on business needs, new jobs being introduced, minor delays, and other everyday occurrences.

Both Oozie and Control-M expose jobs as a service, which simplifies scheduling and management. The main differences between the solutions are in how they support pre-scheduling planning, and the flexibility they provide to make changes after jobs are scheduled. Impetus identified several differences that can significantly impact how schedulers and operations staff can do their jobs. The differences are highlighted in the table below.

Table 2: Key Differences in Managing Workflows

Functionality	Oozie	Control-M
Can a view of dependencies among workloads be provided to help plan scheduling sequence and resource allocation?	No	Yes
Can a specific job within a workload be placed on hold while allowing the rest of the workload to execute?	No	Yes
Can jobs be tagged with different priority levels (e.g., critical, standard, low priority)?	No	Yes
Can paths be created based on job tagging?	No	Yes
Can conditional scheduling/execution (e.g., if/then) be built into jobs?	Basic if/else conditions can be configured	Yes; tagging simplifies the process

Monitoring & Troubleshooting

Ultimately, the reason for all the development and testing is to operate applications in a production environment in the most reliable manner. Management tools must be able to run applications reliably, detect and report errors as quickly as possible, and provide facilities to analyze and correct problems.

Comparing the job monitoring and troubleshooting abilities of Oozie and Control-M is inherently unfair because Oozie is strictly a scheduling tool while Control-M is a comprehensive digital business automation solution that includes scheduling. Nonetheless, Impetus did a limited evaluation to see how well each option met basic needs for the typical job roles of its customers.

Two differences stood out. The first difference was the SLA monitoring and alerting capabilities native to Oozie and Control-M. Oozie can provide email alerts, and developers must create the alerts for conditions that users want to monitor. Oozie also has a web interface that can set up a monitoring system; however, it does not natively support mobile devices. Control-M also has a web interface, email notification, and text notification support. Control-M has native support for monitoring by mobile and includes a desktop monitoring client.

- Control-M for Hadoop provides monitoring and alerting for Hadoop and non-Hadoop workflows, but Oozie can only be used for Hadoop jobs. Control-M can tie the Hadoop and non-Hadoop workflows to business SLAs and monitor the critical path for failures and delays. In the event of a job running longer than usual or a failure in critical path, Control-M can take remedial actions and issue notification alerts to relevant support teams.
- The second major difference is compatibility with ticketing and other management systems. Oozie is essentially a standalone tool that requires development work to integrate with IT management systems. Control-M provides an SNMP interface for easy integration to third-party monitoring and event management systems, and has out-of-the-box integration with the Remedy digital service management platform and with other enterprise event management solutions.

As previously noted, Control-M is compatible with Hadoop and non-Hadoop jobs, and Oozie is Hadoop-specific.

Updating Jobs

Besides scheduling, Oozie and Control-M have different methods for making changes to job properties. Control-M provides extensive find-and-update capabilities that can be used for things like changing job names, such as in the previous example of renaming jobs to promote them to production, or changing the cluster where the job will run. Oozie does not support mass updates, so each job must be reconfigured individually. That was a relatively minor inconvenience in the test exercise, but would have been much more time consuming in a real-world deployment because a significant amount of manual work would be required to change job attributes.

SECURITY

Hadoop environments often contain sensitive data, so they must be secured to meet enterprise standards. More than half (57 percent) of Hadoop users reported experiencing security challenges in their implementations[3]. Impetus highlighted one major challenge in its notes on the Oozie/Control-M evaluation:

> Oozie is lacking on some of the enterprise concerns like security. For example, when a user uses Sqoop action, the complete Sqoop command needs to be provided inline including the plain-text password. It would have been more secure if the password could be masked.

The evaluation by Impetus found three important, practical security differences between Control-M and Oozie:

- Control-M isolates connectivity information for a job from the actual workflow, while Oozie does not. When systems require connection passwords to be updated (a common requirement), isolation allows users to conveniently update the password without having to redevelop/update the workflow.
- Control-M automatically masks passwords; Oozie does not, but can be programmed to do so.
- Control-M is more secure for file transfers because it has native support for encryption, SFTP, and FTP over SSL. Those capabilities must be developed separately for Oozie.

3 Ibid.

CONCLUSION

Control-M for Hadoop, Oozie, Hue, and other open source Hadoop tools each provide the specific functionality needed to schedule Hadoop workflows. **Control-M for Hadoop proved to be 40 percent faster for developing and deploying a Hadoop workflow**. Control-M for Hadoop also has more features to ensure workflows execute without disrupting business activity, such as automatic recovery of failed file transfers, notification and alert capabilities, built-in security, and integration with ticketing systems.

Oozie is an acceptable option for basic, low-volume Hadoop workflow development and scheduling, while Control-M for Hadoop is better suited for managing jobs and workloads that become an essential part of the enterprise environment.

FOR MORE INFORMATION

To learn more about Control-M for Hadoop, please visit **bmc.com/it-solutions/control-m-hadoop**

485448