BEST PRACTICES IN BIG DATA AUTOMATION: THINKING BEYOND OOZIE TO THE ENTERPRISE REQUIREMENTS



For many Hadoop developers, using Oozie would seem natural. It is part of the Hadoop installation, is standard across most Hadoop providers, and presents a way to schedule entire batch processes rather than single Hadoop jobs like Pig and Hive.



Often what many Hadoop developers are missing is how it integrates to the rest of the business, especially if that business is using a batch scheduler or Workload Automation tool such as Control-M or another tool. Some points to consider:

- Most companies already have a group designated for monitoring critical production flows
- Error notifications and alerting are limited in Oozie
- Dependency logic in Oozie is not as robust as enterprise-class Workload Automation
- Hadoop is not a standalone process, it gets data from different sources and sends data back out
- Dynamic processing is limited

Monitoring critical production flows

One of the key differences is who is monitoring the flows. Most companies with Workload Automation have a dedicated team as the first line of defense in monitoring batch. These batch processes are usually considered production flows and contain processes that are critical to the entire business. The production batch environment monitors virtually every application in the enterprise such as Data Warehouse, ERP, Homegrown tools, Supply Chain and many others. This team is the first to see problems or situations that will cause issues for the business. The Workload Automation tools provide UI's for all levels of visibility into the batch flows.

If the flows are defined and run through Oozie, the batch team would have to learn another tool, Oozie, and wouldn't be able to see the end-to-end logic of the entire flow. These flows will instead need to be monitored by the team that created them or as part of the Application Development team. In essence the monitoring of Oozie processes would have to be pushed to a generally more expensive, less centralized team. And most notably, if these flows are critical to the business they are not being monitored with the rest of the processes.

Error notifications and alerting are limited

While Oozie does provide some alerting capability through email based on defining decisions, it is neither robust for other alerting nor simple to define decisions. A Workload Automation tool such as Control-M provides multiple options for either sending alerts or taking actions on alerts. Sending alerts might include SNMP message, email with attached output, or messages to a centralized alert panel. Actions could include rerunning the process, running a self-healing process, opening help desk tickets, posting the dependency for another job to run or taking any of these actions based on a specific error code.

Dependency logic is not as robust

Unlike many tools that include some scheduling function, Oozie gives the developer a way to create flows rather than schedule a single process. It can schedule all of the Hadoop type jobs, but also includes Java, SSH, streaming and shell script jobs. Using shell scripts, the developer can run any command line integration to another third party package. What it cannot do is make direct connections to these packages such as Informatica, IBM Datastage, or other tools.

Workload Automation tools such as Control-M provide true integrations that connect with a third party tool and execute, monitor, and control those processes. It has more robust file watching capability to ensure the data is received before the Hadoop jobs run. It can also run SQL Scripts, Stored Procedures, SSIS and SQL Agent processes against multiple databases to get or post Hadoop data. More importantly it can group both the Hadoop processes and external processes into a single coordinated service. This service can be managed to its Service Level Agreements, sending notifications when it becomes late or troublesome. It can be viewed and monitored through multiple UI's including web and mobile apps.

Dependencies can be created based on multiple scenarios including files, jobs completing or even external processes posting to it. The flows can be time driven or event driven or both. They can run concurrent or single threaded.

Hadoop is not a standalone process

Hadoop developers are concentrating on their Big Data processes and may not be aware of other applications and technologies that IT Operations needs to manage. They consider the process done when their Oozie flows are complete.

In reality, other parts of the enterprise require output from Hadoop in order to do their jobs and update their systems. Those shops adding Hadoop into tools such as Control-M have created services to include these non-Hadoop processes. In many cases multiple jobs and applications need to process data before Hadoop can perform its functions. The output from Hadoop will be fed to other systems as well, all of which are being monitored and controlled by the Workload Automation tool. Since alerting and notification has been set up for these non-Hadoop and Hadoop processes, the entire flow is taken care of.

Dynamic processing is limited

Most Oozie flows are scheduled based on a time or bundled with a time. While the development team has access and can run flows dynamically, the ability for outside teams to do the same is not

there.

Using a Workload Automation tool such as Control-M, any user can be provided the capability to dynamically run, monitor and control their services and jobs. In addition these services are then monitored for being late or having failures that would prevent it from completing in time.

Summary

While Oozie has more capability and options than the traditional scheduling tool integrated into a third party product, it still lacks much of the required enterprise-class functionality upon which the entire business must ultimately rely. Especially compared to functionality available in advanced Workload Automation tools such as Control-M.
Many companies are asking the Hadoop development team to work with the WA teams and create functional flows in the proper tools.

BMC's Complete Guide To Hadoop

1. Introduction to Hadoop ›
2. Hadoop Benefits ›
3. Hadoop Ecosystem ›
4. Hadoop Security ›
5. Hadoop Hiring & Careers ›