A Broadcom Company
technologies

White Paper

# Automic® Workload Automation

## Hadoop Workflow Automation

## Executive Summary

### Challenge

The insight provided by big data has become a prerequisite for companies to remain competitive. As dictated by creating an open source software project, Hadoop has grown from being a small distribution to becoming an integral part of many companies' IT ecosystems. The speed at which a company can generate business insights out of Hadoop is vital to the data they have available when making important business decisions. The challenge is how to integrate new big data applications and processes into existing IT processes without causing major disruption and impacting business-as-usual operations.

### Opportunity

This white paper explores why organizations need a workflow engine to support their big data Hadoop environment. It also highlights the main open source workflow solutions available today and exposes the limitations of these solutions.

### Benefits

The paper then moves on to demonstrate why Automic® Workload Automation is a must-have workflow engine for Hadoop and big data processing. Automic Workload Automation introduces native integration that simplifies and accelerates delivery of enterprise Hadoop applications and insight. It quickly ties Hadoop into larger business processes, exposing it to the broader business user, rather than just the data scientist.

# Why a Workflow Engine Is an Imperative for Hadoop

If you're operating in the big data space, you have an abundance of technologies to consider, making the process of standing up an environment complex. However, when it comes to creating a big data environment, the basic principles of a smaller standard environment are still applicable. You still need to integrate traditional data from your relational data structures into big data systems. And conversely, you need to enable data from those big data systems to be integrated into the traditional environment to produce reports.

The message is clear: When it comes to executing big data workflows, these workflows typically combine big data technologies and legacy applications in a single business process.

So why is a workflow engine an imperative component in a Hadoop development environment? When it comes to data processing, Hadoop developers frequently struggle to process big data in its raw format. Multiple, overlapping and, above all, time-consuming pre-processing operations, such as standard extract, transform, load (ETL), need to take place prior to the physical processing.

To remain agile and meet the timely needs of the business for new services and updates, Hadoop developers need to automate this process by organizing the steps into reusable workflows that can industrialize and accelerate the development process. Put simply, automation eliminates the need to write new code.

# Negotiators versus Workflow Managers

It is important to differentiate between a resource scheduler (often referred to as a "negotiator") and a workflow manager. Negotiators are important components of Hadoop. The scheduler produces processes on different nodes, allocating resources based on application requirements and cluster capacity. Hadoop's Yet-Another-Resource-Negotiator (YARN) operates transparently to the user, and generally you do not have to deal with it.

Workflow managers, meanwhile, manage complex Hadoop tasks—for example, multiple jobs that run sequentially, in parallel or in response to event triggers. A job can be many things, such as running individual Java apps, accessing the Hadoop file system/other data stores or running various Hadoop applications.

The differences don't end there. Hadoop workflow managers are also different in terms of programming model/language, code complexity, property/parameter description format, supported applications, scalability, documentation, and support.

# Open Source Workflow Tools: What's Available?

## Apache Oozie

When a group of Yahoo engineers met around a table in Bangalore, India to find a way to perform more complex, multi-stage Hadoop processing, the result was the Oozie framework. This open source project, based on Java technology, simplifies the creation of workflows and manages coordination among jobs.

Apache Oozie (as it subsequently became known) enables developers to blend multiple jobs sequentially into one logical unit of work. The advantages are abundant. First, Apache Oozie is fully integrated with the Apache Hadoop stack and supports Hadoop jobs for Apache MapReduce, Pig, Hive, and Sqoop. Second, the open source framework can be used to schedule jobs specific to a system, such as Java programs. Third, Hadoop administrators can create complex data transformations that can combine the processing of different individual tasks and even sub-workflows. The result? More control over complex jobs, and increased repeatability of jobs whenever needed.

## Azkaban

Azkaban is an open source workflow engine aimed at the Hadoop ecosystem. Developed by LinkedIn and written in Java, Azkaban resolves the ordering through job dependencies and provides a user-friendly Web interface to maintain and track big data workflows.

There are several common features between Azkaban and Oozie. Both are open source workflow engines for Hadoop job scheduling, and both are written in Java. There, however, the similarities end. Azkaban is simple to use, with easy to define workflow schedules, whereas it is more complex to define workflows using Oozie. Azkaban job scheduling only supports time-based scheduling, while Oozie supports both time-based and input-data-based scheduling. Azkaban keeps the state of all running workflows in memory, but with Oozie, a workflow state is in memory only when doing a state transition.

## Airflow

The accommodation-rental service Airbnb recently opensourced Airflow, its own data workflow management framework, under the Apache license. Airflow is being used internally at Airbnb to build, monitor, and adjust data pipelines. The platform is written in Python, as are the workflows that run on it.

Airflow enables developers of workflows to author, maintain, and run workflows based on a periodic schedule. The platform interacts with Hive, Presto, MySQL, HDFS, Postgres, and S3. Hooks are also provided to make the system more extensible. Airflow provides a command line interface, as well as a Web-based user interface that allows users to visualize pipeline dependencies, monitor progress, and trigger tasks.

How does Airflow differ from Oozie or Azkaban? Airflow pipelines are defined as code, as opposed to a markup language in Oozie or Azkaban. Moreover, tasks are instantiated dynamically, as opposed to creating tasks by deriving classes in Luigi (see the following section). As a result, Airflow is ideally suited to situations where pipelines are generated dynamically from configuration files or metadata of any form.

## Luigi

This is an open source Python-based data framework for building complex data pipelines, and it features built-in support for Hadoop. Created by Spotify in 2009, Luigi is now also used by organizations like Foursquare, Stripe, and Asana in their production environments. Instead of using an XML/YAML configuration of some sort, all the jobs and their dependencies are written as Python programs. It manages multiple tasks, including dependency resolution, workflow management, and visualization.

The Luigi framework is relatively small (consisting of several thousand lines) and is therefore easy to understand. Moreover, the workflow manager provides a straightforward abstraction to define a data pipeline in terms of tasks and targets, and it will manage the dependencies.

Luigi has many benefits: It handles dependencies, cuts the volume of boilerplate code required for error checking, and manages failure recovery. The solution also ensures developers follow a clear pattern during data pipeline development.

## Pinball

Pinterest recently offered Pinball as a scalable open source workflow manager. The solution supports a broad range of data processing pipelines, ranging from jobs composed of simple shell scripts to elaborate Hadoop workloads. Pinball is widely used by Pinterest engineering teams and handles hundreds of workflows that process data on a daily basis in the Hadoop clusters. The largest workflow has 500 jobs.

Out of the box, Pinball supports a Python-based workflow configuration syntax. It also provides a number of job templates for configuring simple shell scripts, as well as more elaborate computations on the Hadoop platform. With Pinball, users have the flexibility to model data dependencies between jobs—for example, by delaying a job until the data it needs becomes available.

# Common Limitations of These Solutions

When it comes to choosing open source workflow solutions such as the ones above, there are several precautionary steps users should consider:

- Each of these open source solutions has been designed by young, fast-growing companies—Airbnb, Spotify, and Pinterest, for example. These organizations are rarely of a type or scale to necessarily integrate big data with legacy production environments. However, most organizations do have requirements to integrate big data with traditional applications like ERP or a data warehouse.
- Each of the open source workflow solutions identified above is oriented toward big data engineers and developers, not data analysts or standard IT users. Today, most typical big data use cases involve nontechnical users like data analysts that require self-service capabilities. This might be to perform agile analytics or other processes.
- Another limitation of these open source solutions concerns the compliance and data governance that are mostly lacking, or need, custom code. As part of a GRC strategy, IT organizations must ensure that the business has trustworthy, compliant, and authoritative data that is secure from inadvertent or malicious loss.

# Too Many Tools = Too Much Management

If you adopt too many tools into a process, the management burden can be overwhelming. For each tool or application you design into your big data workflow, you must find a way to integrate, synchronize and coordinate multiple tasks between it and the multitude of other tools used in the system.

This process is complicated to build and manage. Most tools lack the visibility and intelligence required to understand what is happening in all other parts of the system. Tool selection, integration, and management are long-standing challenges in the computing industry. Many organizations expend significant time, money, and resources exploring which tools they want to use, only to discover that even more work is needed to understand how to use them, integrate them and manage them.

It doesn’t stop there. Once an organization has overcome these challenges, even more time and effort are needed to build the pieces required to integrate the tool assortment. The explosion of newly available big data applications and tools has only accelerated the level of the challenges.

# Why Automic Workload Automation Is a Must-Have Workflow Engine for Hadoop

If you are looking for a Hadoop workflow automation solution that is built to meet the demanding enterprise requirements of a modern-day business, then it’s important to consider the following key capabilities:

- Native Hadoop job scheduling capability
- Operating system and Web service agents (SOAP and REST) to future proof all future Hadoop developments
- Built-in managed file transfers
- Highly available and scalable architecture
- Seamless upgrades, with zero downtime
- Support for object orientation to enable fast, flexible, and reusable workflows
- Native third-party application support, including ERPs and BI

- Support for dynamic, event-driven automation support for advanced workload balancing
- Integrated SLM capabilities showing both business and IT perspective
- Integrated workload analytics and reporting delivering insights on workflow activities
- Self-service capability to enable business users' involvement
- Ability to enable security and compliance across the enterprise
- Long-standing history of automation across all verticals
- Missing zero downtime upgrade capabilities, which is not possible with open source tools

While there are many other factors that are also important to consider, only by ensuring that your solution encompasses the key capabilities listed above as a foundation can you be certain that you will be able to meet the ever-increasing demands of the business.

Automic Workload Automation delivers on all these capabilities. Our Hadoop automation solution introduces native integration that simplifies and accelerates delivery of enterprise Hadoop applications and insight. It quickly ties Hadoop data into the larger business process, exposing it to the broader business user rather than just the data scientist.

# Automic Workload Automation Feature Comparison

| **Tool Focus** | **Enterprise Workflows** | **Hadoop Workflows** | **Hadoop Workflows** | **Python Workflows** | **Python Workflows** | **Python Workflows** |
|---|---|---|---|---|---|---|
| Workflow Triggers | Time, Predecessors, File, Application, Event | Time, Predecessors | Time, Predecessors | Time, Predecessors | Time, Predecessors | Time, Predecessors |
| Hadoop Cross Cluster Dependencies | Yes | No | No | Yes | Yes | Yes |
| Map Reduce/ Yarn | Yes | Yes | Yes | Yes | Yes | Yes |
| Pig and Hive | Yes | Yes | Yes | Yes | Yes | Yes |
| Sqoop | Yes | Yes | Yes | Yes | Yes | Yes |
| High Availability | Native, OOTB | Loadbalancer | No | No | No | No |
| Configuration Interface | GUI and Scripting | XML | Java Property Files | Python Code | Python Code | Python Code |
| Target User | Non-Technical to Technical | Technical | Technical | Developer | Developer | Developer |
| Graphical UI | Dev, Monitoring, Admin, Ops, Service Catalog | Monitoring | Monitoring, Ops, Admin | Monitoring, Admin | Monitoring | Monitoring |
| File Transfers | Native | No | No | Need Code | Need Code | Need Code |
| OS Support | Wide (15+) | Narrow (Linux) | Narrow (Linux) | Python | Python | Python |
| Application Integrations | Wide (SAP, Oracle, Informatica, and so on.) | No | No | No | No | No |
| ITSM Integration | Yes | No | No | No | No | No |
| Version Management of Workflows | Yes | No | No | No | No | No |

| Tool Focus | Enterprise Workflows | Hadoop Workflows | Hadoop Workflows | Python Workflows | Python Workflows | Python Workflows |
|---|---|---|---|---|---|---|
| Self Service GUI with User Prompts | Yes | No | No | No | No | No |
| Pre and Post Conditions for Each Job | Yes | No | No | Need Code | Need Code | Need Code |
| Workflow Restart Capability | Yes | No | Yes | Need Code | Need Code | Need Code |
| Workflow Rollback Capability | Yes | No | No | No | No | No |
| Zero Downtime Upgrade | Yes | No | No | No | No | No |
| Data Governance and Compliance | Native | No | No | Need Code | Need Code | Need Code |
| Integrated SLM | Yes | No | No | No | No | No |
| Integrated Analytics | Yes | No | No | No | No | No |

For more information, please visit ca.com/automation.