Databand CPD Workflows integration

Detecting anomalies in IBM Cloud Pak for Data workflows using IBM Databand

Thomas Schwarz schwarzt@de.ibm.com
Sergej Schütz sersch@de.ibm.com
Md Intekhab Shaukat Md.Intekhab.Shaukat@ibm.com

Summary

This developer code pattern demonstrates how to surface governance artifact workflows of IBM Cloud Pak for Data and their user tasks in IBM Databand. This leverages the broad alerting and analysis capabilities of IBM Databand to detect anomalies in the workflow progress and enables alerts for cases when tasks take too long. That way, a workflow supervisor can monitor workflow operations with ease and get alerted on potentially stuck workflows.

Description

IBM Knowledge Catalog is a data governance software that provides a data catalog to automate data discovery, data quality management, data lineage and data protection. It is available as managed SaaS and within IBM Cloud Pak® for Data. It provides all the required means to share trustworthy and validated data in a data marketplace. In order to protect governance artifacts like business terms, data classes, classifications, reference data sets, governance rules, and governance policies, changes to such artifacts need to run through an approval workflow. Such workflows can be very simple like automatically publishing without any approval up to involving multiple approval steps and even client-specific custom flows. In IBM Cloud Pak for Data, a workflow supervisor has only basic capabilities to monitor workflows and detecting those that require special attention can be a tedious task. The integration with IBM Databand allows the supervisor to leverage all its capabilities to monitor and analyze the progress of all running workflow instances using the dashboard as well as detailed workflow instance views. In addition, setting up duration alerts allows for a timely detection of any unattended workflows using the multitude of alerting mechanisms supported by IBM Databand.

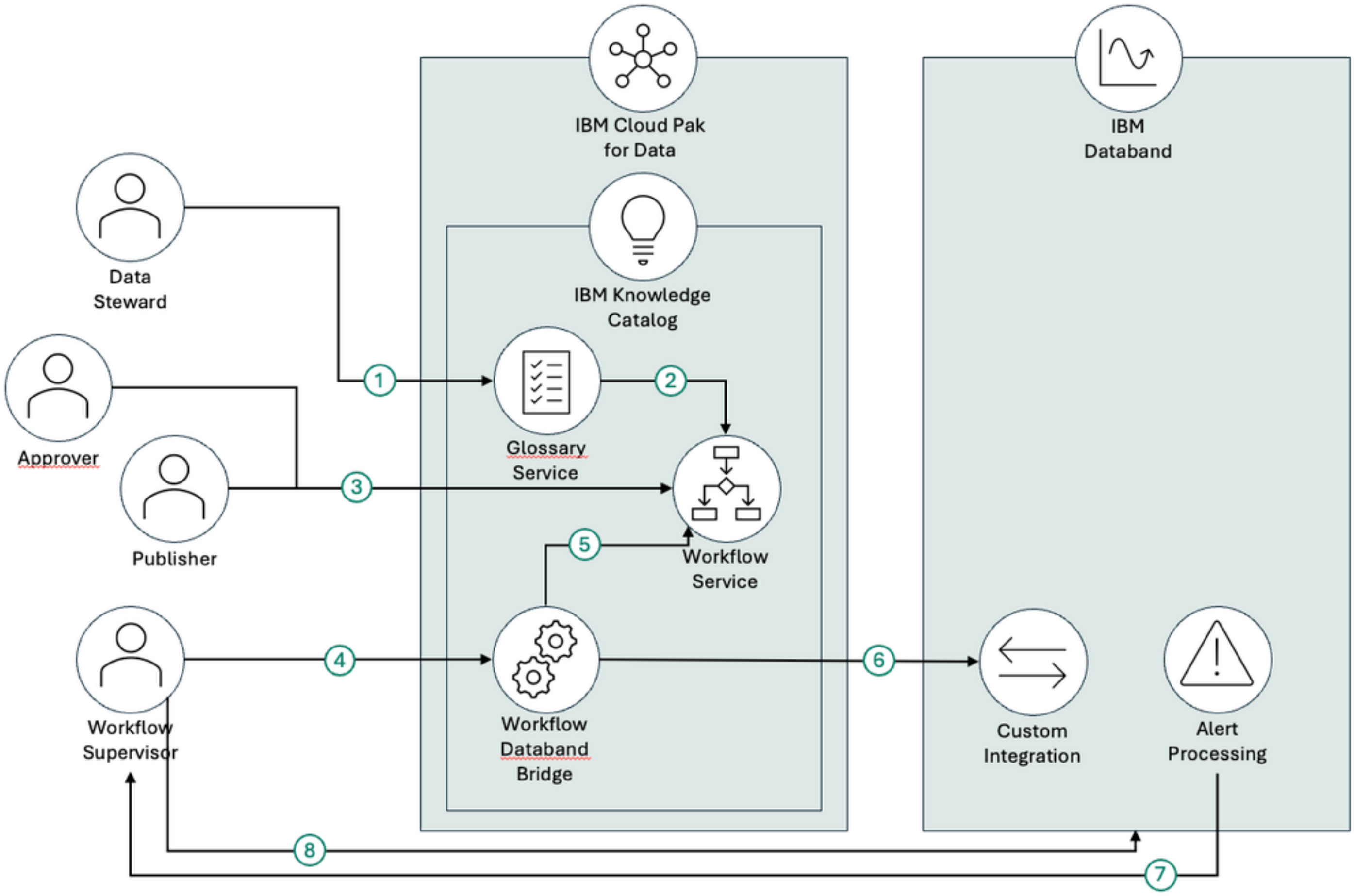
In IBM Knowledge Catalog, which is offered as part of IBM Cloud Pak for Data as well as on IBM Cloud, data stewards work with the governance artifacts defined in the Business Glossary. The glossary requires regular updates and additions, e.g. in order to meet new regulatory requirements or to setup new data quality goals. Each artifact change results in the creation of a draft artifact as well as a workflow instance that implements the governance process for the change. A variety of different workflows with differing numbers of branches and steps gets launched depending on the configuration, with the specific workflow template typically determined by the artifact type and its primary category (see here for details). The configuration defines which people need to collaborate to perform steps like reviews and approvals until finally, the draft gets published and the change becomes available to all Business Glossary users.

Sometimes, workflow tasks can get left unfinished for long periods of time or even indefinitely, e.g. when users become sick, take time off or switch jobs. If the corresponding workflows are configured such that the absent user is the only assignee on a task, or if the user *claimed* the task as their own, no other users will see the task in their task inbox, resulting in a situation where an open task goes unnoticed by the whole team. Using an integrated IBM Databand setup, the workflow supervisor can remedy such issues in a timely fashion by taking advantage of automatic alerts that get triggered when workflows take longer than expected. Once an issue is detected, the supervisor can either *unclaim* the task or assign other users to it so that processing can continue.

In IBM Databand, each workflow type such as *governance artifact workflows*, *data quality remediation workflows*, or *potential suspect workflows* get represented as a separate *project*. Workflow configurations that define specifics like when to launch a workflow, which flow of tasks to use and which users to assign are represented as *pipelines* in IBM Databand, allowing the workflow supervisor to define separate alerts on each of them. Workflow instances get represented as *job runs* and user tasks are mapped to *tasks*.

For the sake of simplicity, the Workflow Databand bridge in the presented code pattern takes the IDs of single governance artifacts as its input and submits the current state of the corresponding workflow to IBM Databand. The bridge code can be run periodically to sync the
progress of the workflow to IBM Databand.

Flow



Regular user interaction

1. IKC data steward introduces a change to a governance artifact in the Glossary
2. Glossary service creates a draft for the change and launches a workflow
3. Various users like Approvers and Publishers collaborate on the workflow, and eventually the change gets published or discarded

Workflow supervisor interaction

4. The Workflow Databand bridge is (periodically) run to sync workflow progress to IBM Databand
5. Workflow Databand bridge collects workflow status and user tasks statuses
6. Workflow Databand bridge transforms IKC objects into pipeline runs and tasks and sends to custom integration API in IBM Databand
7. IBM Databand sends out alerts as configured, e.g., run duration alerts
8. Workflow Supervisor monitors workflow operations using the Databand Dashboard

Instructions

Setup instructions

1. Get the sample code from the “databand-integration” folder in the https://github.com/IBM/wkc-workflow/tree/main GIT repository, and navigate into your local copy of the “databand-integration” folder.
2. Create a file “config.properties” as described in README.md
3. In IBM Databand, go to the Integrations section, Click “Add integration”, click “Custom integration”, give it a name, and then copy the URL and token to the DATABAND_URL and DATABAND_TOKEN properties in the config.properties file
4. add your IKC connection details to the same file as well.

IKC usage instructions

1. In IBM Knowledge Catalog, perform any change on any published governance artifact, e.g., create a new business term. Under the covers, this will launch a new workflow instance to control the approval of the draft.
2. work with the draft and the user tasks of the corresponding workflow in IKC
3. copy the artifact_id/version_id from the draft’s URL in your browser. The ids are the 2 guids (globally unique identifiers) at the end of url.

Databand monitoring instructions

1. Launch the Workflow Databand Bridge and provide the artifact_id/version_id as command line parameter.
2. In IBM Databand, setup a Pipeline duration alert on the workflow pipeline
3. Continue using IKC and the bridge, and see how alerts get triggered, and how the dashboard evolves.