IKC data quality integration with IBM Databand

Building a monitored data quality pipeline with IBM Knowledge Catalog and IBM Databand

Sergej Schuetz sersch@de.ibm.com
Mike Grasselt grasselt@de.ibm.com

Summary

This code pattern implements a deep integration between the IBM Knowledge Catalog data quality capabilities and Databand. By exposing individual IKC data quality calculation steps as a Databand pipeline, users can leverage Databand's powerful monitoring and alerting features to catch data quality issues.

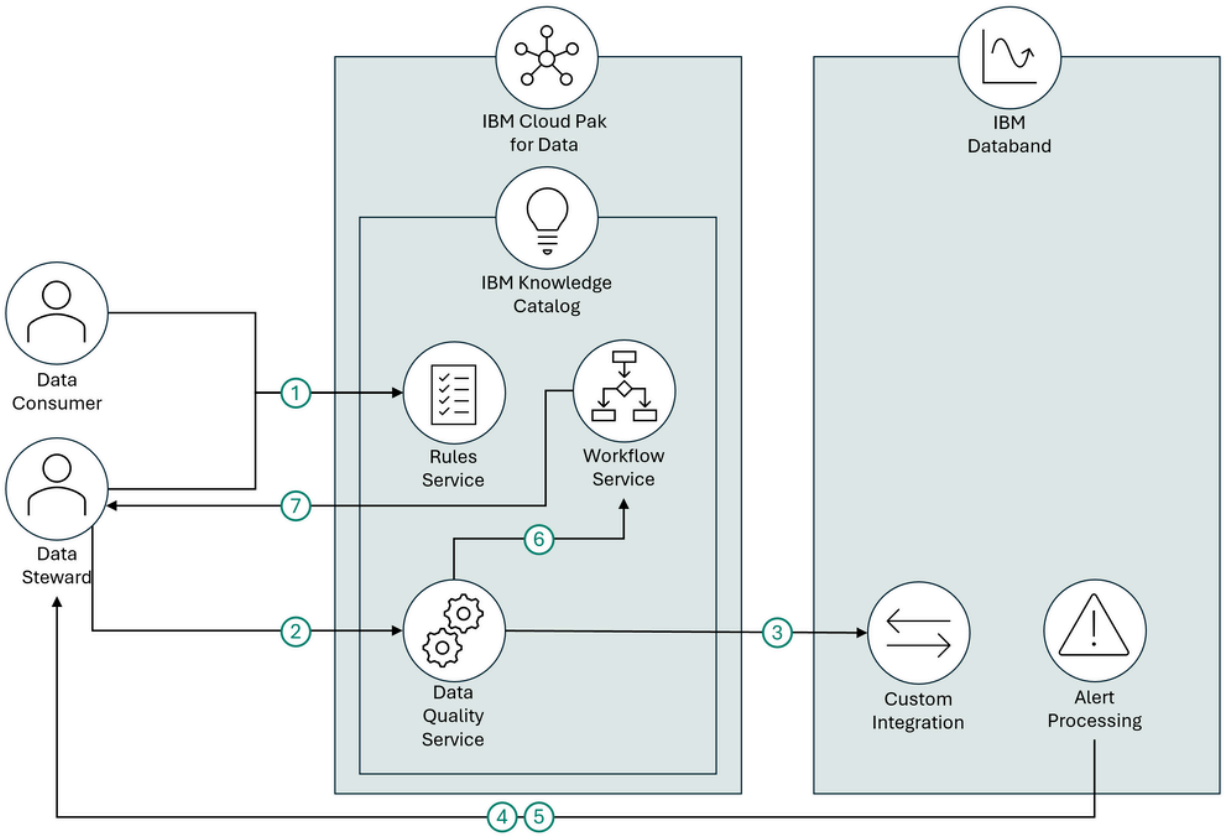
Description

IBM Knowledge Catalog is a data governance software that provides a data catalog and automates 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. A core capability is the automated data quality process for monitoring and remediating the quality of data assets with scalable data quality service level agreements (DQSLAs). IBM® Databand® is observability software for data pipelines and warehouses that automatically collects metadata to build historical baselines, detect anomalies and triage alerts to remediate data quality issues.

While analyzing data assets, the data quality process in IBM Knowledge Catalog computes scores for various data quality dimensions such as accuracy, completeness, or consistency. For example, a table column might have a low completeness score if many records have missing values in that column. All column scores are combined into asset scores, grouped by quality dimension. These dimension scores are then summarized into an overall data quality score for the asset. This helps data steward teams quickly identify data assets in the data lakehouse that need to be corrected and notify the responsible data stewards. To automate this process, the data quality process can automatically compare the scores against the acceptable thresholds defined as part of customizable DQSLAs. For example, a DQSLA for critical customer data assets might require an overall asset data quality score of at least 98%. If a data asset falls below this threshold, the DQSLA is considered violated. For this situation, the DQSLA specifies a remediation workflow that will automatically trigger a task for the responsible data steward.

The presented code pattern sends reports of the complete asset score computation and DQSLA assessment pipeline described above to IBM Databand, which automatically monitors that pipeline end-to-end. It can trigger custom alerts for data quality degradation in any analysis or score computation step. Furthermore, alerts can be triggered in case the data quality process encounters unexpected issues, which can then be used together with DevOps tools to report the impact of application failures.

Flow



1. Data Steward and Data Consumer create agreed data quality service level agreements (DQSLAs)
2. Data Steward schedules automated data quality process
3. Each data quality process run for an asset is reported and appears as a pipeline run in IBM Databand
4. For each pipeline run, a critical failure (e.g. unexpected decrease of data quality scores over time) triggers a notification (slack or email) of the responsible data stewards for remediation
5. For each pipeline, a run frequency anomaly triggers a notification (slack or email) of the responsible data steward
6. If the asset violates the DQSLA, a workflow is triggered
7. The data steward will see tasks created by the workflow in the task inbox

Instructions

1. In IBM Knowledge Catalog, create a project and import a data file, e.g., a CSV file, as described here.
2. Create a metadata enrichment job that runs predefined data quality checks for the imported data asset as described here.

3. Optionally, create data quality rules as described here.
4. Review the asset and dimension scores with the asset data quality page as described here.
5. Create SLA rules for the asset score and the dimension scores as documented here.
6. Rerun the metadata enrichment job to trigger the analysis and SLA assessment.
7. Run the Databand integration provided with this code pattern to report the data quality pipeline data to IBM Databand.
8. In IBM Databand, evaluate the run and create alerts for the pipeline and pipeline tasks.