DTSQUARED

Integration techniques between Collibra and Cloudera

Version – Iteration 1.0

info@DTSQUARED.co.uk 02070304975

DTSQUARED Ltd: 60 New Broad Street, London, EC2M 1JJ

REVISIONS

0.1	CS – Original
1.0	TP - Review

TABLE OF CONTENTS

INTEGRATING COLLIBRA AND CLOUDERA

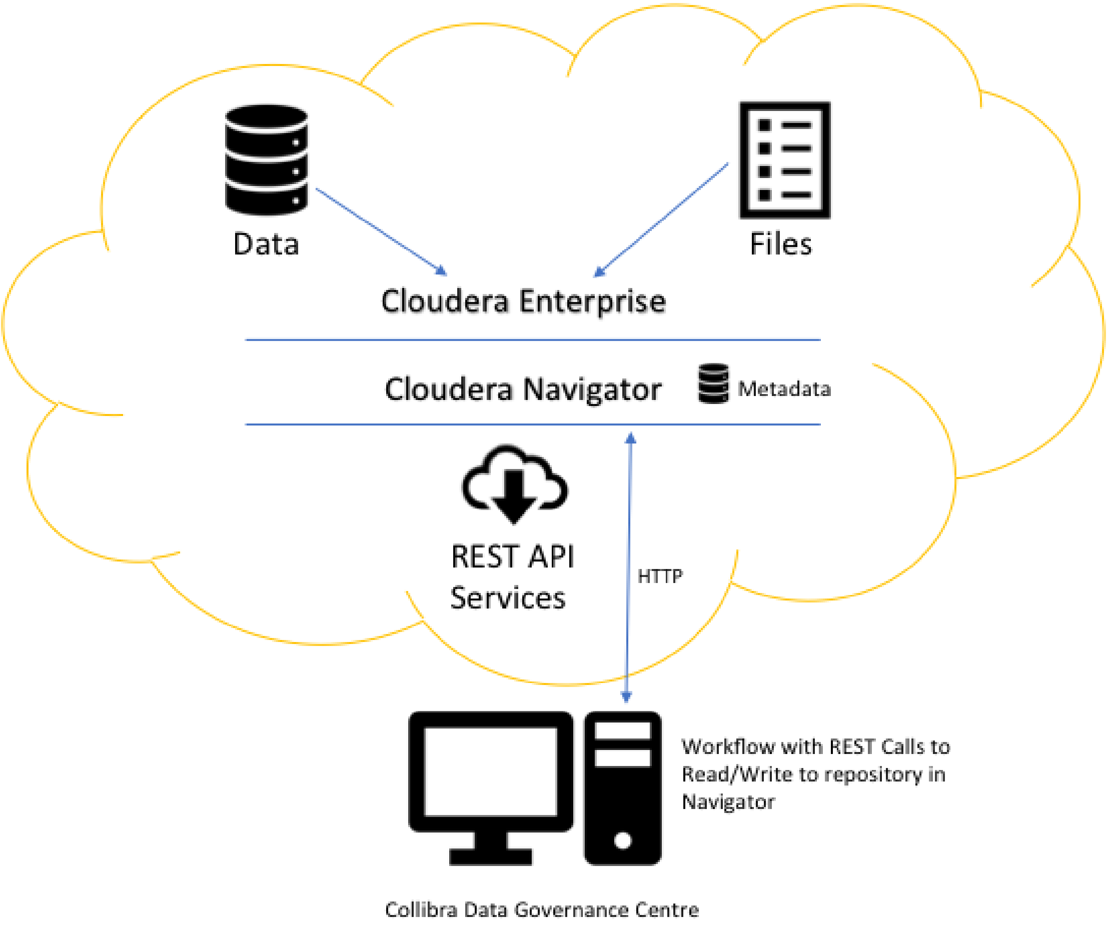
OVERVIEW

This document is a view into the integration of Collibra and Cloudera, specifically for Assets, which could be controlled through Collibra and linked to the physical asset in Cloudera.

The purpose is to show that Integration is possible and how this may be achieved.

SCOPE

The scope is to investigate and show some detail around what is possible and some potential scenarios to give this context. This is an iterative piece of research with more detail enabling a customer specific approach and plan.


Data
Files
Cloudera Enterprise
Cloudera Navigator
Metadata
REST API Services
HTTP
Workflow with REST Calls to Read/Write to repository in Navigator
Collibra Data Governance Centre

CLOUDERA AND WHAT WE WANT TO DO WITH IT

Cloudera Enterprise Hub is an implementation of a cloud based scalable environment. Utilising the management tool called Navigator, which enables metrics via custom metadata tagging and queries across assets within the cloud.

Navigator is a web based UI which we can report against the assets within the cloud and is integrated to using the REST API which will enable read and write functionality to the Cloudera metadata.

PRE-REQUISITES - ASSUMPTIONS

This document has been prepared with some pre-requisites and required configuration against the specific environment, as detailed below: (Obviously if these are not as described then we would work with you to get to this position)

- There is a running physical implementation, which utilises control and metrics that we want to use (data, SQL, file,) under a Cloudera enterprise deployment, typically Hadoop.
- All the services are running in the implementation that we are intending to utilise & integrate with navigator (HDFS, hive, SQL etc.).
- However if functionality exists within un-supported service, we write via REST API entries to have one source (this requires further definition by the client).
- Cloudera Navigator (and the API Port setup) are configured and available on the network
    - Configured meaning that the required metadata metrics are being logged and match the needs (out of the box there is a set of metrics). This needs to be aligned to process, and should be at v5.0 and running given it supports custom metadata, a required feature to enable this type of integration.

INTEGRATION

To set the context for this document, we need to explain what integration means.

Integration meaning first linking assets in Collibra with Assets in Cloudera:

- a licence file of a system
- a field in a database
- a table, Surname as a Data Element of a Customer Data Element
- etc.

The Integration needs to link the assets in both Collibra, where control may be enforced, to the physical asset and its state being queried.

Once linked there is a need to maintain the State of the Asset post Collibra integration.

WHAT ARE WE INTEGRATING

Here are a few examples of specific Collibra controlled needs to address (when we run a workflow against a Collibra asset) we may want to:

- Test a constraint.
- Expiry policy of data or file
- Security policy on access of an asset
- Data or file retention policy

Through custom code, via workflows (step into the Collibra workflow with groovy code to make a REST call via HTTP, make available Cloudera metadata at point of script, and handle the response), apply policy, constraint and rules against assets in Cloudera that are known about in Collibra.

EXAMPLE SCENARIO

One scenario, could be the retention of an insurance document; we could enforce the upper limit of: a metadata attribute; a date field attached to an asset of type insurance document; and upon change make a rest call to Navigator to enquire as to the availability of the asset in Cloudera (Consume Cloudera Metrics).

If an asset`s expiry date has not passed / in the event of the asset being unavailable, we could update the Collibra Asset to reflect this and we could even attach custom metadata to the asset in Cloudera to persist the status of the integration if required (Augment Cloudera metrics).

***Something to be careful of:** A deleted file will exist in Navigator metadata, if the date range of the stored metrics covers it, but if we purged metrics in Navigator, these may not be available.*

Issues and errors can be handled by persisting at asset level: what problem(s) have occurred and we can get access to Reports via views in Collibra, post changes, could show the 'Issues' born out of the workflow execution.

CONSIDERATIONS

THINGS TO PLAN FOR IN CLOUDERA NAVIGATOR

There are issues around Navigator that we must factor in around the amount of data that is captured at metadata level, especially with very active installation, this is reliant on what we want and how long we need it for? (purging and archiving of the data usually in line with compliance policy within an organisation). Persistence of metrics past the lifetime of an asset would be beneficial to the clarity of feature functionality, **this needs to be understood.**

THINGS TO PLAN FOR IN COLLIBRA

Change process, how this will be achieved. There are a number of solutions to this which we will not cover here.

INFORMATION AND ACTIVITIES REQUIRED WHEN WE BEGIN

- What is the level of integration, detail, scope?
- Availability of metrics in Cloudera that we need, **versions of installation**, **capabilities**.
- **How we identify links between assets**, what's common and unique to drive categorisation?
- **Naming conventions around duplicates and constraints within Cloudera and Collibra need to be understood and aligned as a starting point**
- **How we manage change** (frequency, deletions, new items, duplicate items that may break uniqueness?)
  - Change in Collibra first (this we can manage, but what are the expected events we will be listening for (file change, table updated, expiry missed)
  - Change in Cloudera first? (someone deletes a file from an ftp server for example), what is expectation? if known.

FINDINGS

Integration is a readily available capability of both Collibra and Cloudera, but more definition around the Cloudera implementation is required, to start to solidify and understand what deliverables and capability is available and configured to the enable the agreed process and as such there is always a discovery phase to the integration piece.