# PDP-8: Integrated Data Management

## Motivation

PNDA today has a number of data management components, some of which are integrated together, while others are are complementary tools.

- Data Service exposes a REST API and manages policy for archive or deletion on the basis of dataset age or size
- A daemon we call 'hdfs cleaner' looks after cluster housekeeping by moving & deleting unwanted artifacts that otherwise accumulate on the system
- The Console has a Data Management page which exposes Data Service functionality
- Compaction can be configured to compact old datasets using Gobblin in order to lessen 'small files' issues in some deployments
- HDFS has the capability to designate types of storage for specific data, for example it's possible to move old data to cheaper spinning disks and use SSDs for active data

However these are somewhat disjoint and require operator intervention to build end to end data management policies that use all these capabilities together.

What we would like is the ability to simplify this process for the PNDA operator and bring these and other data management capabilities together so they can be closely coordinated & audited.

## Alternatives

Rather than building in this area, there are existing projects that need to be evaluated. Apache Falcon and Apache Atlas are two of these.