PDP-7: Securing PNDA

PNDA forum discussion

PNDA-4031 - Security: authentication, authorization & perimeter IN PROGRESS

Motivation

This is a large area and is divided into several separate topics -

Access control between users & services

Identity - a consistent notion of what constitutes an identity across PNDA & the execution of functions for that identity

- E.g. applications started by Bob are then associated with Bob for the purposes of resource allocation

Authentication - a consistent approach to establishing the veracity of identities across PNDA

- E.g. Alice cannot pretend to be Bob

Authorization - a consistent approach to controlling what an authenticated identity can & cannot do across PNDA

- E.g. Bob can configure ingest and create applications, whereas Alice can only control the lifecycle of applications that have already been created.

Some of the key areas to be addressed include -

- Most PNDA services are identity-aware. However, there's no consistent authentication of identity.
    - Where PAM is the underlying framework authentication is deferred to the configured mechanism, which could be local or LDAP
    - Some services can be configured to use LDAP directly
    - Other services assume a default identity without any authentication
- Today, all services are accessed directly. Access control is greatly simplified by having one control point through which all services are accessed.
- Some PNDA specific services do not currently implement authorization
- For the services that do implement authorization, there are a multitude of schemes and control mechanisms.
    - Management of authorization is greatly simplified by having a consistent approach, ideally with a single point of management.

Securing interaction between services

This also sub-divides into identity, authentication & authorization. Typically, we will use TLS on links and mutually authenticate on certificates.

Proposal

- Introduce identity to PNDA services where missing today
- As far as possible, introduce one point through which access to PNDA is controlled
    - We believe Apache Knox is the most suitable technology for PNDA in this space, having a wide range of applicability across Hadoop services, pluggability to support PNDA services and supporting a number of widely used authentication frameworks out of the box.
    - Some services are not covered by Knox and need separate analysis of how to provide a consistent authentication scheme overall.
    - More about Knox.
- Authenticate identity consistently across PNDA services
    - Kerberos is the key Hadoop technology of interest for strong authentication, but there are other options and other PNDA technologies to consider, as well as applicability to cloud based deployment.
- Authorize operations consistently across PNDA services
    - The key technology of interest here for Hadoop is Apache Ranger.
    - There are services across PNDA that would not be addressed by Ranger and need separate analysis of how to provide a consistent authorization scheme overall.

Overall Plan

Identity

- Add identity awareness to all PNDA services and APIs where missing
- This work is mostly complete

Gateway

- Integrate upstream Knox version 1.0.x as part of PNDA deployment in order to cover main HDP and other services.

Authentication

- Authentication at the perimeter using Knox is pluggable - LDAP is likely, OAuth needs some investigation
- Authentication behind the perimeter likely to be based on Kerberos & some components will need work to enable this

Authorization

- Initially, distributed management of authorization & a simple fixed scheme based on users (not roles)
    - See here for how this will be achieved for the Deployment Manager in the short term
- Later, centralized authorization, likely using Ranger.

Phases

- **Phase one - basic perimeter, authentication & authorization**