PDP-9: Support for Druid

PNDA forum discussion

PNDA-4519 - Integration of Apache Druid in PNDA as OLAP solution for real-time event data IN PROGRESS

Motivation

- Apache Druid is an open-source data store designed for sub-second queries on real-time and historical data. It is primarily used for business intelligence (OLAP) queries on event data.
- Druid provides low latency (real-time) data ingestion, flexible data exploration, and fast data aggregation. Existing Druid deployments have scaled to trillions of events and petabytes of data.
- Druid is most commonly used to power user-facing analytic applications. It can load both streaming and batch data and integrates with Samza, Kafka, Storm, Spark, Flink and Hadoop.
- Druid can be considered as a OLAP support option from the previously proposed Hadoop based OLAP tool Kylin in PDP-4.

Proposal

- Provide the Druid UI as part of PNDA and integrate it with the PNDA components such as the PNDA Console and the PNDA Deployment Manager. Druid support is optional per deployment.

- Druid Cluster Diagram

Design

The following section discusses the changes required to each PNDA component.

PNDA Mirror

Druid resources and any other dependencies will be hosted on the PNDA mirror. The mirror build script will need to include these in the appropriate mirror section.

Druid Components in PNDA

For Druid cluster, will launch new nodes for Druid borker, Historicals and MiddleManagers, Coordinator and Overlord processes and use the existing nodes of PNDA cluster for kafka, zookeeper, and mysql for druid metadata storage.

Support will be added for deploying and configuring Druid components in heat templates and salt configuration files respectively.

Deployment Manager:

A Druid component plugin will be created that will run druid applications. A supervisor will be set up on the PNDA edge node that will call the druid CLI to process the durid query operation.

Console:

The PNDA console dashboard page will be modified to include add Druid blocks under data storage.

Logging

Each druid component will have a specific log file for debugging purpose.

Example applications

The community druid example applications will be created that demonstrates use of druid.

PNDA Guide

Sections of guide will need creating or updating to reference Druid

Plan

Phase 1 - Integrate of Druid with single node deployment using Openstack Pico flavor.

(Refer http://druid.io/docs/0.12.1/tutorials/quickstart.html for Druid single node deployment.)

- Along with the changes made from the above 6 components and corresponding documentation effort. the following tasks will be fulfilled:

    1. Data ingestion through kafka/Tranquility
    2. Data ingestion status display in Druid console from PNDA console
    3. Sample OLAP queries from REST client
    4. (stretch goal) The above can be verified in AWS Pico setup with the help from community.

Phase 2 - Integration of Druid cluster in Openstack Standard flavor & AWS Standard flavor

- Same to phase 1, but extended to these 2 flavors.

Phase 3 - Druid stand alone (lambda integration) vs server cluster deployment

- Start Druid or set up connection to standing Druid cluster at PNDA creation.
- OLAP queries to Druid data from PNDA console.
- Support Druid Health monitoring at PNDA console.

- Druid cluster interfacing directly with Kafka.

Notes:

- Tranquility could be installed along with Druid as the real time event data injection mechanism consuming data from the data/message bus.

Interfaces

- Expose the Druid native APIs.
- Integrate with Spark and Flink as stretch goals from phase 2 (need more discussion).

Compatibility

- TBD

Alternatives

- Need to study and document further upon whether or not Druid and Kylin can be deployed, configured and running along with each other.