Portable Hadoop Environment

# Mission Statement

The mission of the Portable Hadoop Environment (PHE) is to provide data professionals with a way to rapidly provision a barebones hadoop environment that allows the importing, manipulating and dissecting of data all from any machine with Docker installed.

# Description

The PHE has been designed to be a minimalistic environment to focus on the core aspects of the Hadoop ecosystem while still resembling a realistic architecture. While there are intentions of adding more components through a modular container design, the essentials will always be the center of focus. HDFS, YARN, Hive, Spark represent the core components of the PHE.

While the current iteration uses Hadoop and Hive 2, future development of the PHE will include Hadoop and Hive 3 as its own independent environment.

# Uses for the PHE:

* Training
* Development
* Proof of concept
* Continuous Development or Continuous Integration

# Maintainers

Tim Bytnar - [tim.bytnar@daugherty.com](mailto:tim.bytnar@daugherty.com)

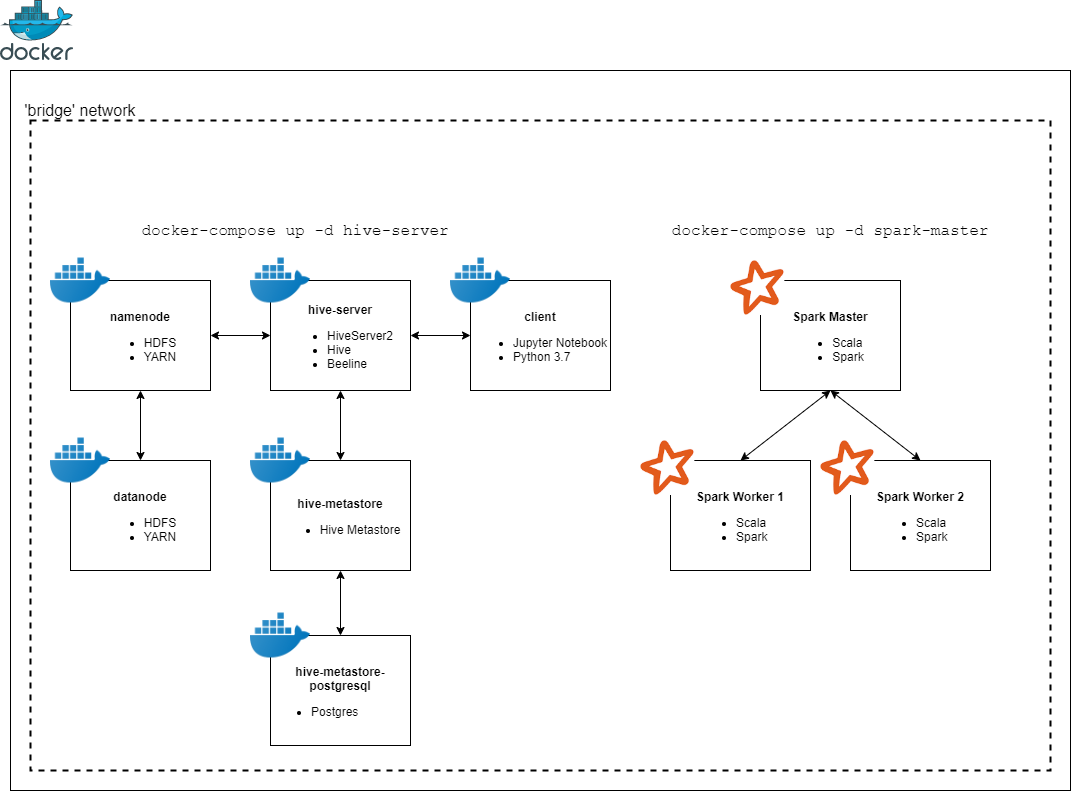
# Container Architecture

The Hive/Hadoop environment consists of six containers:

* Namenode - The primary (master) Hadoop container
* Datanode - The secondary (worker) Hadoop container
* Hive-server - HiveServer2 and all Hive executables
* Hive-metastore - Hive Metastore
* Hive-metastore-postgresql - PostgreSQL (Metastore Database)
* Client - Jupyter Notebook and Python

The Spark environment consists of three containers:

* Spark Master
* Spark Worker 1
* Spark Worker 2



# Future Plans

1. Expand the client containers capabilities
2. Add more Hadoop ecosystem components as modular containers
3. Improve Cross Platform Compatibility and Reliability
4. Baseline Performance “Edition”
5. Baseline Security “Edition”