Running Application using Docker Container on HDP 3.1

Presenter : Tejhan Bharadwaj Ramkumar
University of Maryland Baltimore County

Mentor : Todd Larchuk

Coordinator : Steve Polston

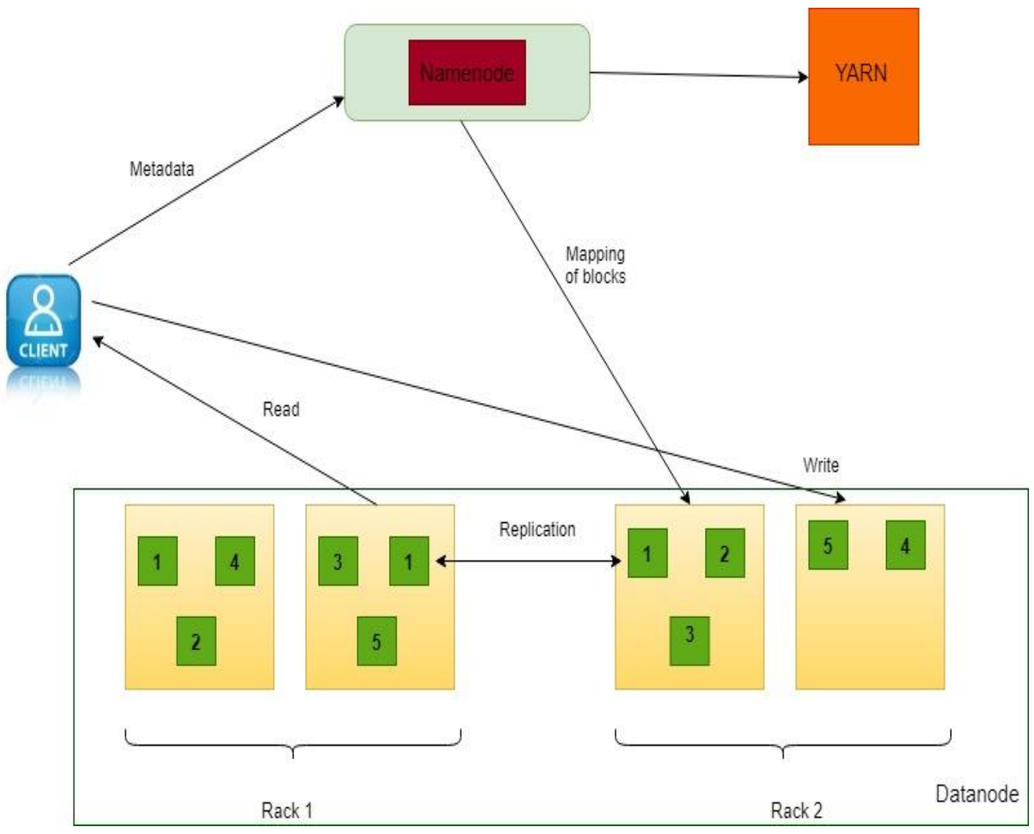
Introduction

- The main objective of this project is to deploy Hadoop 3.1 cluster with Docker installed on all the working nodes, then configure the cluster properly in such a way that it supports both yarn and Docker containers and demonstrate that it works by running some applications.

Design

HDFS Architecture

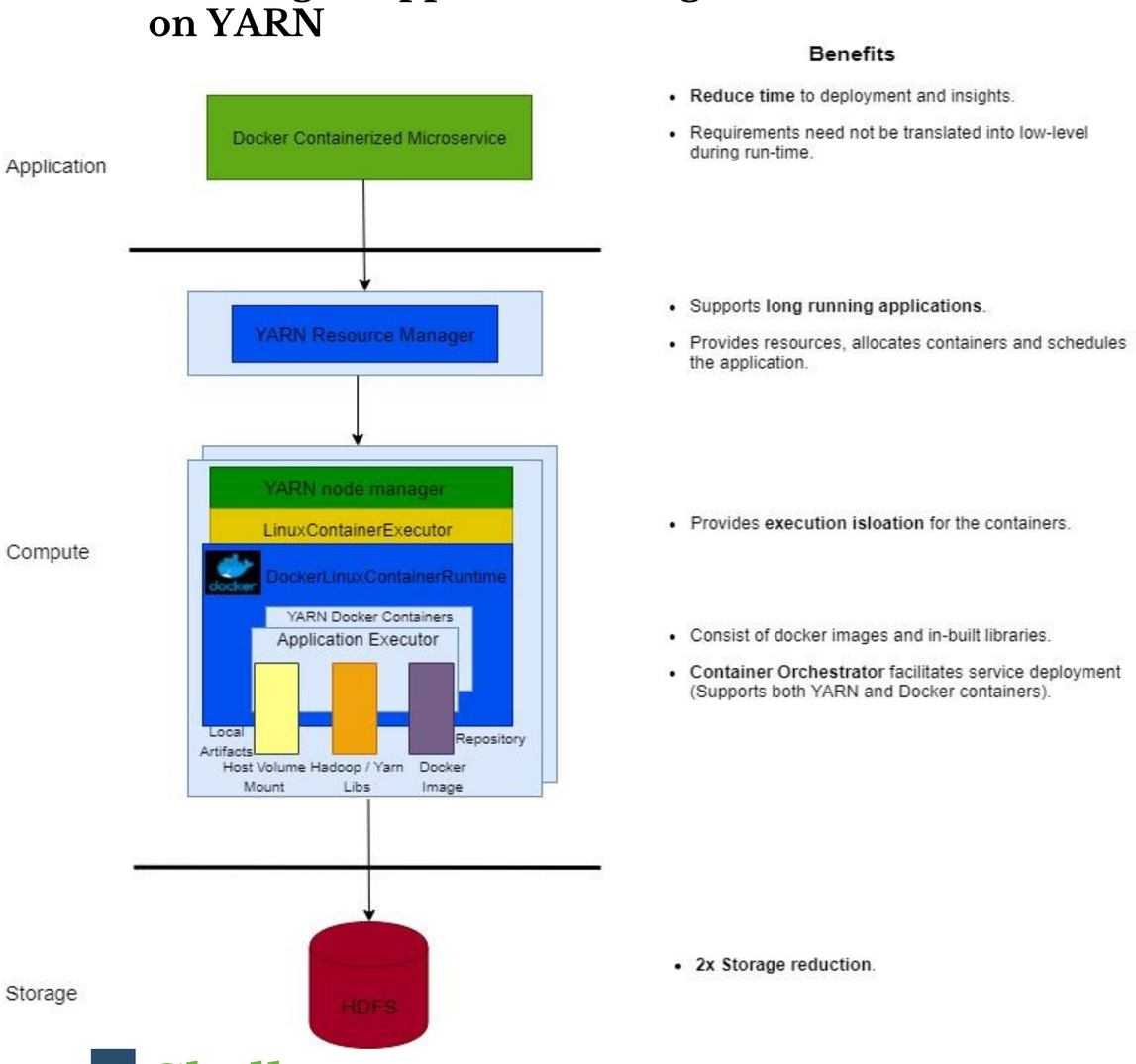
- **Hadoop Distributed File System.**
- **This is our storage architecture.**



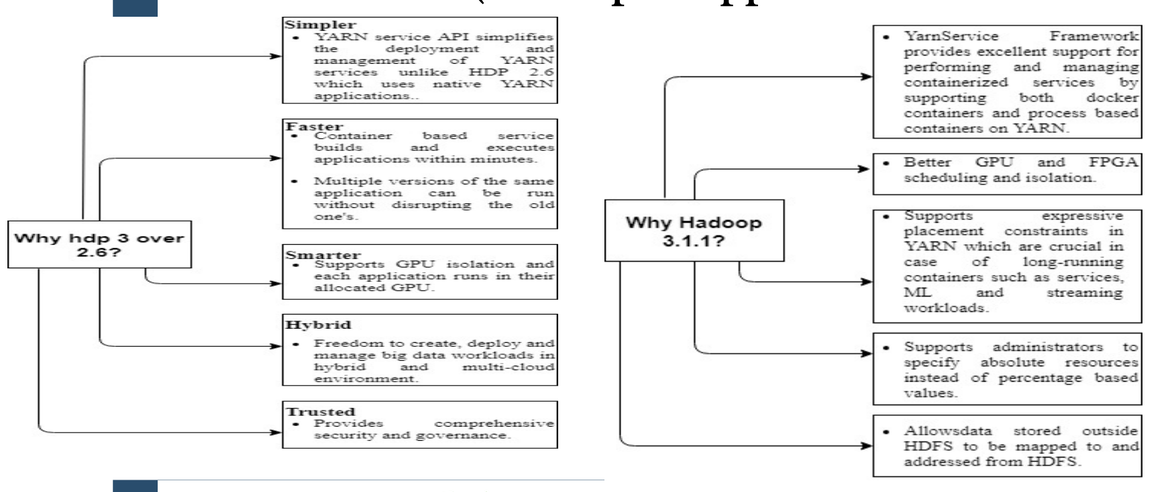
- Master/Slave architecture which consist of namenode, datanode and client.
- Client send the metadata to the namenode and can also read and write the files.
- Datanodes manage the storage and usually stores the data as files in one or more blocks.
- Datanodes also manages the creation, deletion and replication of blocks upon request from namenode.
- Namenode manages the file system namespace operations like opening, closing and renaming. It is also responsible for mapping of blocks in the datanodes.

Phase –1 Model

Running an application using Docker containers on YARN

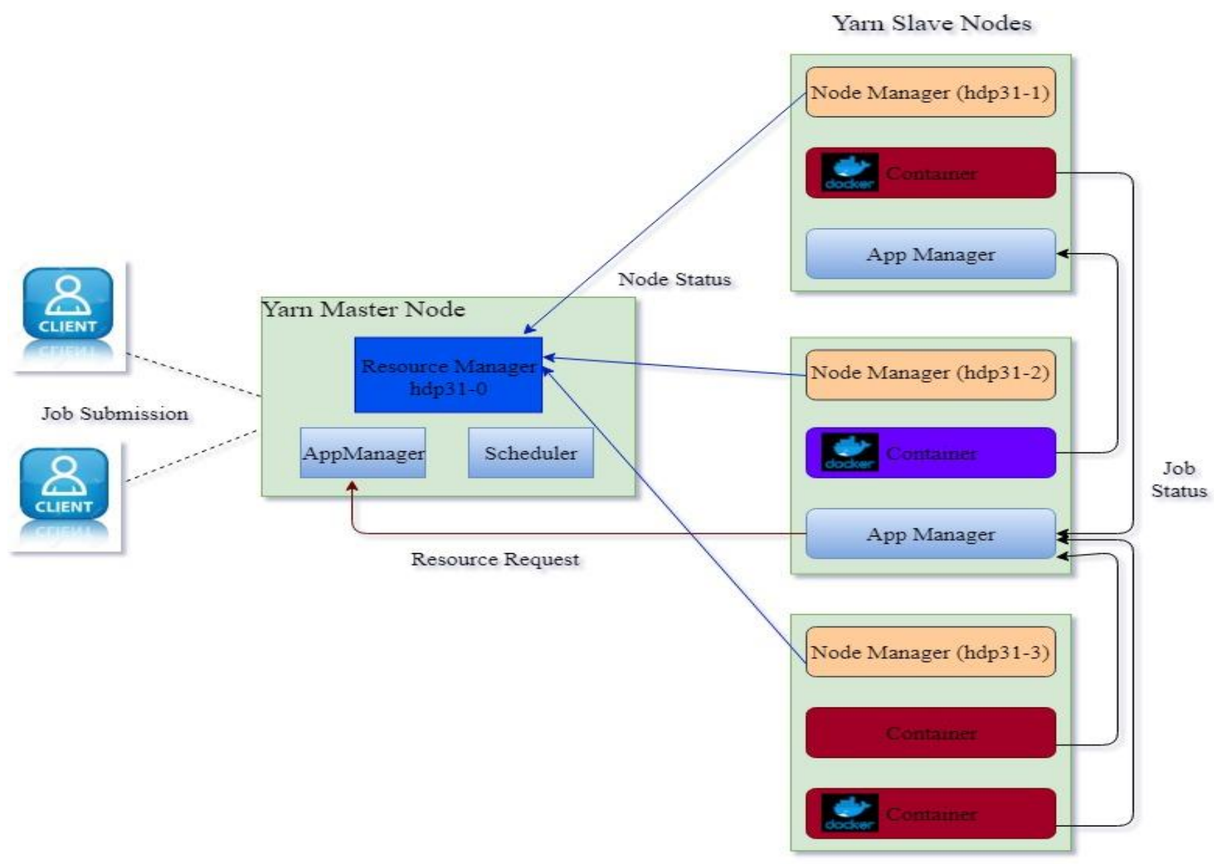


Motivation (Hadoop 3 supports Docker containers)



YARN Architecture

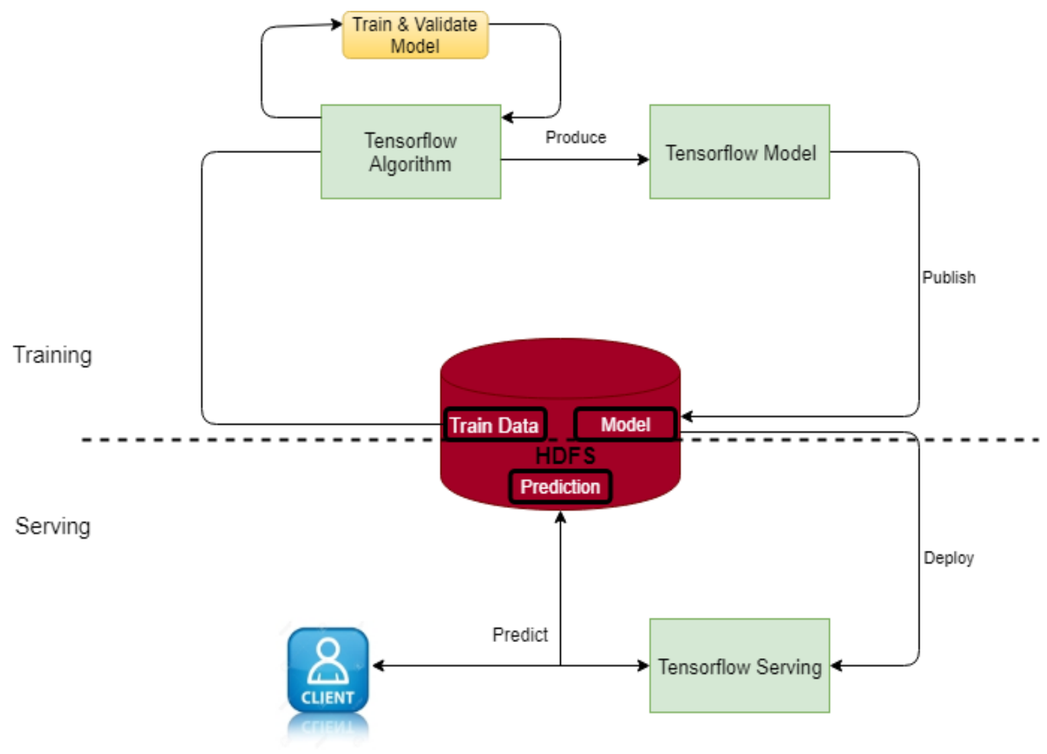
- **Yet Another Resource Negotiator.**
- **It is a framework for running the applications.**



- Resource manager and node manager form the data computation framework.
- Resource manager is the ultimate authority and it arbitrates resource among all the applications in the system. It has two major components: Application Manager and Scheduler.
- The Application Manager is responsible for accepting job-submissions, negotiating the first container for executing the application specific Application Master and provides the service for restarting the Application Master container on failure.
- The Scheduler is responsible for allocating resources to the various running applications based on the resource requirement of the application.
- Node manager is a framework agent who is responsible for the containers, monitoring and reporting the resource usage and the status of the node it is present in.

Phase –2 Model

Running a machine learning model using Docker containers on YARN



Challenges

- ✓ Application Json file trying to access the file inside the Docker container during execution.
- X Manually placing the Docker image on all the VM's, since we don't have any pre-defined method to do this.
- ✓ Read/Write data from/to respectively from HDFS into the Docker container.

Future Works

- If Hadoop 3.1 cluster with Docker container is configured properly and deployed on all the working nodes is able to execute a Machine learning algorithm then in future this could pave way for ACUMOS to deploy dockerized Machine learning models into the Hadoop cluster.