# A SURVEY ON AUTOMATED DEPLOYMENT OF CLOUDERA DISTRIBUTION ON DOCKER CONTAINERS

[1]*SHRIKANT S. RAUT,* [2]*SWATI SALEM,* [3]*RUPALI KALOKAR,* [4]*SUNIL NAGARGOJE*

[1, 2, 3, 4] **Department of Computer Engineering, Sinhgad Institute of Technology, Lonavala, Savitribai Phule Pune University, Pune**

***shrikantr976@gmail.com, swaatirao@yahoo.co.in, rupalikalokar@gmail.com, sunilnagargoje00@gmail.com***

***ABSTRACT :*** *With the increase in data day by day in several terabytes, it seems nearly impossible to manage and process such large data repositories. So there has to be some technique to process data efficiently. Increased use of virtualization had made it possible to run several different distributed applications on same hardware with maximum hardware utilization and flexibility. These days such virtualization facilities are provided by various virtual machines, but when it comes to virtualization resource overhead is the major factor which is to be considered. Traditional virtualization technologies fails to manage the overhead thus comes into picture the linux containers that prove to be a powerful alternative with less overhead, more flexibility, maximum resource utilization and high performance. In this paper we surveyed on present virtualization technologies and their alternatives which can provide efficient and enhanced use of existing hardware resources and easy packaging and transportation. We have explored on main advantages of docker containers and how data processing can be made easier and faster by automating deployment of hadoop on these containers using appropriate automation techniques..*

***KEY WORDS :*** *Repositories, Virtualization, Overhead, Docker, Containers, Hadoop, Multitenancy.*

## 1. Introduction

In the world of ever increasing data (Bigdata), there are few solutions (frameworks) to manage this large amount of data. Also for efficient use of existing hardware, there are various virtualization technologies including traditional virtual machines that prove to be a better solution for hardware utilization. In order to analyze and process large amount of data existing data analysis and processing frameworks (such as hadoop) are being used with the virtualization technologies to provide efficiency and increase hardware utilization up to certain extent. Virtual machines use the concept of hypervisor for monitoring different guest Operating Systems. The guest Operating System in virtual machine also runs with separate kernel, thus with host Operating System's kernel every guest Operating System will have separate kernel, this results in increasing overhead and increased startup time. Hypervisor also provides isolation between different guests Operating Systems, but due to separate kernel the process isolation results to be very expensive. Thus traditional virtual machines lack in minimizing the resource overhead, more start up time and are not able to tolerate faults.

Thus docker technology is a better solution over traditional Virtual Machines so as to analyze and process large amount of data with easy transportation in production using containerization support.

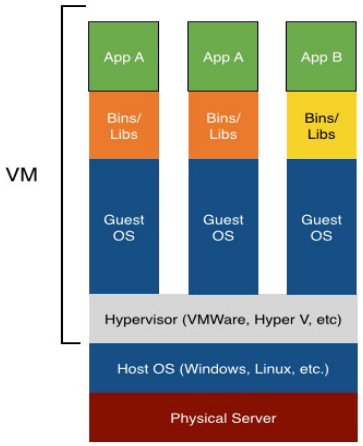


Fig. 1. Traditional Virtual Machine

They do not use hypervisor approach and provide support to containerization approach. Wherein all the guest Operating Systems run on top of same host Operating System using host machine's kernel. In this case each Operating System run as separate processes maintaining same isolation that traditional virtual machine provides. [4]

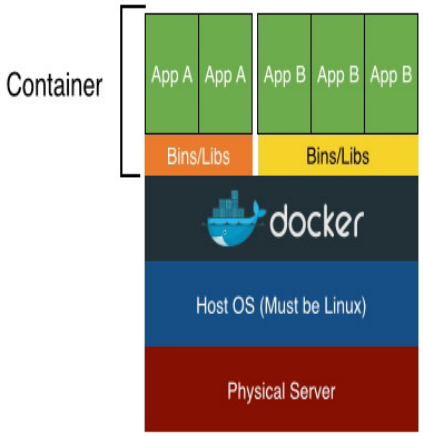


Fig. 2. Containerization Approach

Bigdata in general refers to large data repositories that contain huge volumes of data and it is not efficient to use traditional data management tools and techniques to process and manage such variable data because this data is a collection of structured as well as unstructured data. There are several frameworks like hadoop to manage Bigdata but there has to be some methods to adderss issue of hadoop deployment which is a slow and complex process. Thus concept of DevOps can be used in order to accelerate the deployment work flow.

## 2. Related Work

Bigdata being a very hot topic,is also an active research area in several recent years, there are many projects and research papers focusing on it. There is a research based on migration of a cloud service by use of docker containers to manage separate engines for each tenant, which made possible a quick delivery of an elastic and multitenant cloud service. [1]

Containerization based Hadoop configuration optimization work proposed by Rui Zhang [2], is broadly applicable to cloud environment's analytics framework to support use of Hadoop in docker driven clouds.

Efficient Prototyping of Fault Tolerant Map-Reduce Applications with Docker-Hadoop [3] is one such paper, that explained a fault tolerant model by prototyping distributed systems which was considered as a difficult task. They explored on use of container based technology for developing a prototyping environment for MapReduce applications, using docker-hadoop they were able to simulate several failure senarios to validate a fault tolerant version of Hadoop.

Performance comparision between Linux Containers and traditional Virtual Machines [4] shows macro benchmarks based on the comparisons. Nearly every aspect of both hypervisor and containerization technology related to performance is considered.

Container Orchestration for Scientific Workflows [5] introduced Skyport, an extension to AWE/Shock, that uses Docker container technology to orchestrate and automate the deployment of individual work flow tasks onto the worker machines. The installation of software in independent execution environments for each task reduces complexity and offers an elegant solution to the installation problems.

In Architectural Solution for Virtualized Processing of Big Earth Data [6], Docker as an virtualization technology was put to a test to observe the processing overhead that virtualization containers introduce and docker succeeded in reducing data transfer cost and showed improvements in processing time.

PaaS vendors have used containers as a means for hosting apps. The various container implementations such as Linux Containers, Docker, Warden Container, lmctfy and OpenVZ have been explored in this paper. The paper also shows how each of these handle Process, FileSystem and Namespace isolation.Some unique features of the above mentioned containers have also been discussed and the way in which they differ from the Linux base container. In the end there are few features that are missing in the existing implementations and can be implemented for the next generation PaaS. [7]

Virtualization technology is one of the important part of cloud computing system which is now emerging as a leading edge in computer technology. Virtualization is nothing but creating a virtual version of something which could be an operating system, storage device, or computer network resources or a virtual hardware computer platform. Docker is the latest type of virtualization technology and its applications and advantages in cloud computing are been described. [8]

## 3. Proposed System

We have seen different techniques such as hadoop for managing big data with various approaches such as virtualization and containerization. These techniques are good enough to manage the data but setup(installation) is quiet complex for an ordinary man and takes time while settingup also there are some draw back when using hadoop with traditional virtual machines. We have also seen the advantages of docker based linux containers over traditional virtual machines with the use of containerization in gaining support to obtain multitenancy. Thus we proposed to build a system that helps in automating deployment(installation) of Cloudera's Hadoop distribution on docker monitored linux containers. This automation process makes use of CHEF (i.e. CHEF server, CHEF workstation) as an intermediate to set up hadoop nodes based on users information using CHEF recipes. The users information about the cluster can be obtained via user

interface and can be futher passed on to CHEF workstation using appropriate API, this information can then be used to construct CHEF recipes. CHEF server and workstation then handels the further deloyment using these recipies based on obtained information. The use of CHEF makes the whole automation process look simple with smooth flow and docker monitored linux containers help in decreasing the performance overhead and easy transportation of the nodes.

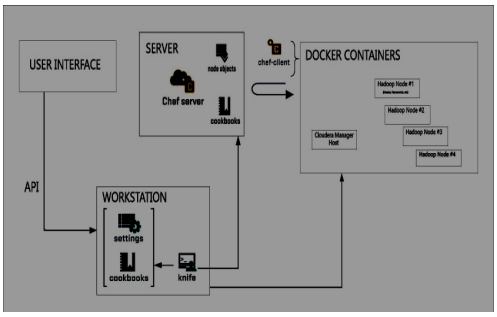


Fig. 3. Proposed Architecture

**4. Conclusion**

From above survey and research work we have reached to a conclusion that, Docker and its container workflow can pack, share, and deploy applications in production environments. CHEF an automation framework that works well with Docker by using a powerful interface to manage various containers, when used to automate the deployment of clouderas hadoop distribution on docker containers makes the workflow smoother and easier. Thus as modern applications/frameworks such as hadoop involve complex deployment pipeline, use CHEF with docker containers provides best solution to above discussed problem.

**5.Acknowledgement**

We would like to express our special thanks to Persistent Systems Pvt. Ltd. who gave us the golden opportunity to do such wonderful work, which helped us to do a lot of research and increase our knowledge.

**6. References**

[1] Aleksander Slominski, Vinod Muthusamy, and Rania Khalaf, ”Building a multi-tenant cloud service from legacy code with Docker containers”, IEEE International Conference on Cloud Engineering, 978-1-4799- 8218-9/15 2015 IEEE

[2] Rui Zhang, Min Li and Dean Hildebrand, ”Finding the Big Data Sweet Spot: Towards Automatically Recommending Configurations for Hadoop Clusters on Docker Containers”, 2015 IEEE International Conference on Cloud Engineering, 978-1-4799-8218-9/15 2015 IEEE

[3] Javier Rey, Matias Cogorno, Sergio Nesmachnow, Luiz Angelo Steffenel, ”Efficient Prototyping of Fault Tolerant Map-Reduce Applications with Docker-Hadoop”, 2015 IEEE International Conference on Cloud Engineering, 978-1-4799-8218-9/15 2015 IEEE

[4] Prof. Ann Mary Joy, ”Performance Comparison Between Linux Containers and Virtual Machines”, 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA), 978-1-4673- 6911-4/15 2015 IEEE

[5] Wolfgang Gerlach, Wei Tang, Andreas Wilke, Dan Olson, and Folker Meyer, ”Container Orchestration for Scientific Workflows”, 2015 IEEE International Conference on Cloud Engineering, 978-1-4799-8218- 9/15 2015 IEEE

[6] Mihai Bica, Victor Bacu, Danut Mihon, Dorian Gorgan, ”Architectural
Solution for Virtualized Processing of Big Earth Data”, 978-1-4799- 6569-4/14 2014 IEEE

[7] Rajdeep Dua, A Reddy Raja, Dharmesh Kakadia, ” Virtualization vs Containerization to support PaaS”, 2014 IEEE International Conference on Cloud Engineering, 978-1-4799-3766-0/14 2014 IEEE

[8] Diliu, Libin Zhao, “The Research and Implementation of Cloud Computing Platform based on Docker”, 978-1-4799-7208-1/14/$31.00 ©2014 IEEE