Hadoop as a Service in OpenStack

Shivaraj Kengond, D. G. Narayan and Mohammed Moin Mulla

Abstract Over the years, data generated by the social media is extremely large in volume. This large volume of data is known as big data. Several physical resources are required to store big data; in order to avoid, these cloud resources can be used. Cloud resources are provided by Amazon Web Services, OpenStack, Rackspace, and many more. Stored big data can be used for analysis of traffic management, call center optimization, real-time fraud detection, social media, and sentimental analysis. Solutions for analyzing big data on cloud are presently not available. Users must obtain the essential cloud computing resources and install the necessary software manually. This can be a cumbersome task for complex distributed services. To address this issue, the services should be viewed as a single application consisting of virtual machines. Users should no longer be concerned about individual machines or their internal organization. To overcome this problem, an OpenStack cloud system with multi-node setup to provide Hadoop as a service to the users has been implemented. Users can easily access the Hadoop service which is provided by OpenStack cloud environment.

Keywords Cloud computing · Hadoop · OpenStack · Openstacksdk

S. Kengond (✉) · D. G. Narayan · M. M. Mulla
School of Computer Science & Enginering, KLE Technological University, Hubli, Karnataka, India
e-mail: shivaraj.kengond@gmail.com

D. G. Narayan
e-mail: narayan_dg@bvb.edu

M. M. Mulla
e-mail: moinbvb@gmail.com

V. Sridhar et al. (eds.), *Emerging Research in Electronics, Computer Science and Technology*, Lecture Notes in Electrical Engineering 545,
https://doi.org/10.1007/978-981-13-5802-9_21

1 Introduction

Big data analytics is an important aspect of business in decision and strategy making. Big data as the word indicates its huge volume of data such as business records, collected real-time sensor data which are used in Internet of things. Data can be of structured (e.g., SQL database) and unstructured data (e.g., files). In order to process big data, Hadoop framework can be used. Hadoop is an open-source project where it provides framework for big data analysis. It works on MapReduce and Hadoop Distributed File System (HDFS) techniques. HDFS is designed to store very large amount of data. To provide Hadoop service and store large amount of data, cloud is needed. Cloud consists of pool of resources, and these resources are accessed via Internet, for example, OpenStack and Amazon Web Service [1]. OpenStack is an open-source cloud platform, which is used to store big data in order to provide Hadoop service to the users.

Figure 1 represents the block diagram of Hadoop as a service in OpenStack. Hadoop Web site provides different Hadoop cluster size to the user. Once the user requests for Hadoop cluster through Hadoop Web site, Hadoop Web site internally calls the script to launch virtual machines and installs Hadoop cluster on launched virtual machines through scripts. Once the installation of Hadoop cluster on virtual machines is done, OpenStack private cloud in response to user's request provides Master IP and Private Key. By using Master IP and Private Key, user can access the Hadoop cluster.

2 Literature Survey

Big data demands new techniques and paradigms for processing data. Hadoop is one of the frameworks which can be used for processing of big data, which is fast, scalable, and reliable. Working of Hadoop and necessity of Hadoop service motivate to carry out the literature survey.

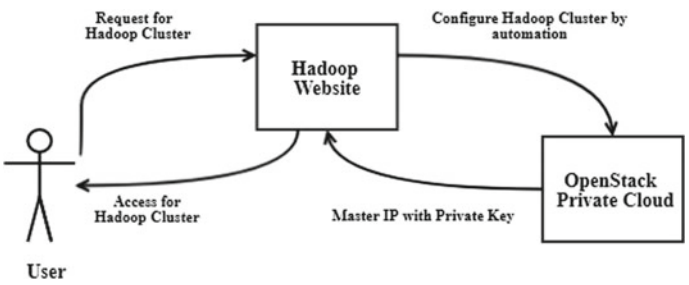

Fig. 1 Block diagram of Hadoop as a service in OpenStack

Authors in [2] present an overview of service-generated big data and big data as a service. In later part of paper, authors explain about requirements to provide big data as a service such as cloud and big data technologies by considering different types of service generated, and data has been discussed.

Authors in [3] discuss the architecture of Hadoop and briefly explain about components of Hadoop such as MapReduce, HDFS, YARN, and common utilities. Hadoop's scalability and fast nature have been highlighted by considering book circulation data along with online profile user data, which consists of 13 GB.

Authors in [4] describe the architecture of OpenStack and storage as a service on cloud using OpenStack. Further discusses about the security aspect of stored data by using AES algorithm.

Authors in [5] explain taxonomy of cloud computing services and different type of cloud services. The taxonomy provides classification of current and future cloud computing services in the form of tree structure.

2.1 OpenStack

OpenStack is an open-source cloud project. It is used for building private and public cloud. It provides three types of services, i.e., infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS).

Infrastructure as a Service (*IaaS*): Cloud infrastructure provides services to customer by virtualizing storage, operating system, server, and hardware. Virtual machine hosting on cloud comes under server virtualization where users host virtual machines on cloud infrastructure depending on their requirements. Users can deploy their own software and manage the virtual machines, e.g., Amazon Web Service (AWS), Microsoft Azure, OpenStack, and many more [5].

Platform as a Service (*PaaS*): Cloud service provider provides platform to user to deploy applications. Users can build and deploy applications using tools and programming languages which are provided by the cloud service provider. Common services provided are Web server to deploy Web applications and different programming language IDE's such as C, C++, Python, and many more. Users have no control over cloud infrastructure such as storage, network, operating system, and servers, e.g., Google App Engine, Amazon Web Service (AWS), Microsoft Azure, OpenStack, and many more [5].

Software as a Service (*SaaS*): Cloud service provider hosts the software and makes this software available to users over Internet. Software as a service removes the burden on users to install and run software on their own machines. Users have no control over the infrastructure such as storage, network, operating system, servers, and software platform, e.g., e-mail service, Google applications, and many more [5] (Fig. 2).

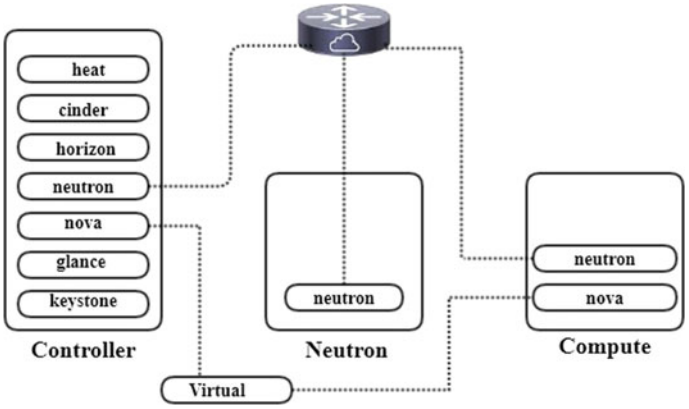

Fig. 2 Architecture of OpenStack

2.1.1 Components of OpenStack

(a) *Keystone*: It provides authentication services for the users and maintains session state using tokens.
(b) *Glance*: It provides storage to store images, which are used to launch virtual machines such as Ubuntu, Cirros, and many more.
(c) *Nova*: It provides virtual machines and management module to drivers, which interact with hypervisor to provide virtualization.
(d) *Neutron*: It provides APIs to cloud tenants to build network topologies and configure advanced networking policies such as switches and routers.
(e) *Horizon*: It provides graphical user interface (GUI) to users to use all the services such as nova, glance, and keystone.
(f) *Cinder*: It provides block volume to virtual machines. Volumes are like external hard drive to the virtual machines.
(g) *Heat*: It provides templates to automate human activities such as launching of an instance, creating flavors, creating networks, and many more.

2.1.2 Multi-node OpenStack

Multi-node OpenStack consists of three nodes namely controller, compute, and neutron node.

(1) *Controller node*: It provides APIs and scheduling of virtual machines and controls the resources of OpenStack such as database, message queue, identity service, and image services.
(2) *Network node*: It provides network service by virtually creating public and private network for virtual machines and network to access the virtual machines externally.

(3) *Compute node*: Compute node runs the hypervisor on which the virtual machines are installed and provides resources to virtual machines such as RAM.

2.2 Hadoop

Hadoop is an open-source framework developed by Apache, and it is based on Java. It is used to store and analyze large amount of data. Hadoop is being used by many companies such as LinkedIn, Twitter, Facebook, Google, Yahoo, and many more, due to its scalable and offline processing nature [3] (Fig. 3).

2.2.1 Components of Hadoop

(a) *MapReduce*: It is a software to process large amount of data in parallel. It consists of two tasks, i.e., map and reduce. Map function is used to tokenize the input data and convert tokenized data as key/value pair. Reduce function takes input from map function and combines the key/value pairs to make smaller key/value pairs.
(b) *Hadoop Distributed File System* (*HDFS*): It is developed on the basis of Google File System. HDFS breaks the data into blocks and stores the blocks over distributed nodes.
(c) *Yet Another Resource Negotiator* (*YARN*): YARN manages the resources and clusters. It is used for graph processing, interactive processing, and many more.
(d) *Common Utilities*: It provides Java libraries and utilities to start the Hadoop module in order to work with Hadoop file system.

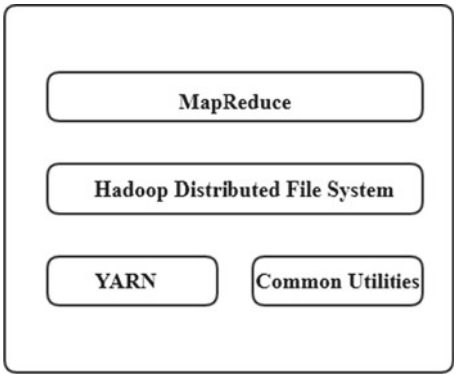

Fig. 3 Architecture of Hadoop

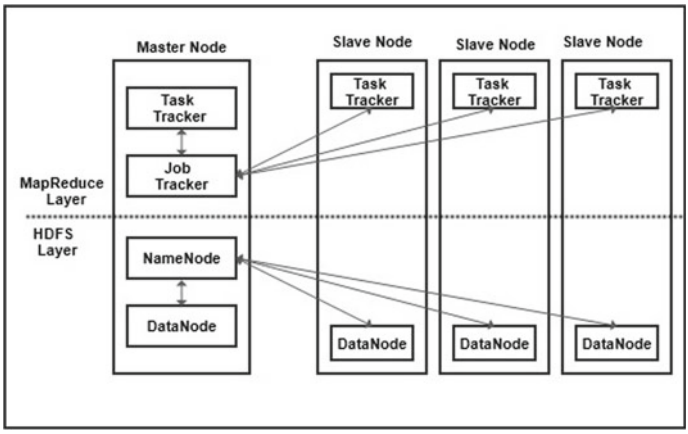

Fig. 4 Hadoop cluster

2.2.2 Hadoop Cluster

Figure 4 represents the Hadoop Cluster. It consists of master and slave nodes. Hadoop works on MapReduce and HDFS. MapReduce consists of TaskTracker and JobTracker, where JobTracker coordinates the tasks in parallel and TaskTracker checks the state of task of each node. HDFS consists of Namenode and DataNode, where Namenode maintains the look-up table for DataNode and DataNode stores the data to process.

Hadoop cluster consists of five daemon process, and all daemon processes have to be up all the time in order work on Hadoop cluster. Master node consists of Namenode, JobTracker, and secondary Namenode. Namenode maintains the lookup table for the Datanodes which resided on slave node. JobTracker monitors the task tracker of all the individual Datanodes and submits back all job status to client. Secondary Namenode maintains the checkpoint of Namenode metadata as shown in Fig. 4.

3 Proposed Work

Hadoop is one of the platforms to process big data. Hadoop can be configured on computers. However, it is difficult to process huge data due to limited storage and RAM. In order to process huge data, pool of resources are needed which can be provided by OpenStack private cloud as shown in Fig. 5. Openstacksdk provides OpenStack APIs to access the resource of OpenStack. Proposed system provides

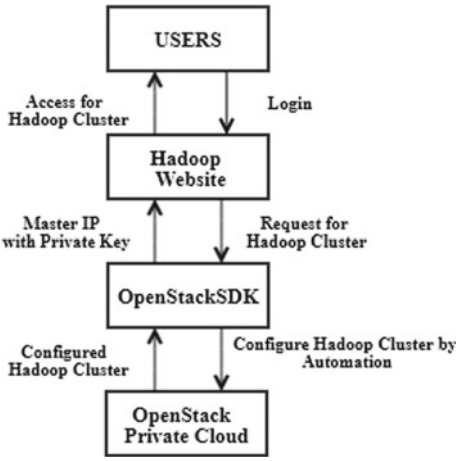

Fig. 5 Proposed system

Hadoop as a service to the user which is configured by automation. Users can access the Hadoop service directly by using IP address and Private Key of the master node.

4 Implementation

This section shows the deployment of openstack private cloud and Hadoop on openstack cloud.

4.1 Deploying OpenStack Private Cloud

OpenStack cloud is scalable and easily deployable to provide storage and network. Cloud is deployed on a single physical server (Intel Xenon ES-2620). Server has 32 GB RAM and 2 TB storage. Cloud consists of four nodes, i.e., 1 controller, 1 neutron, and 2 compute nodes. Deployed cloud is the Newton version of OpenStack. Virtual machines are used to build cloud on server. OpenStack components such as dashboard, keystone, glance, neutron, swift, and nova are configured using protocols and software bundles [5, 6].

(a) *Network time protocol* (*NTP*): It is used for time synchronization between OpenStack private cloud nodes such as controller, neutron, and compute. As OpenStack cloud is distributed, maintaining global time between the nodes is essential for proper functioning.
(b) *Maria DB*: Most of the distributed operating systems require databases to store data. These databases are accessed by all the components of OpenStack (glance, cinder, nova, keystone, and neutron). In OpenStack private cloud, the entire database resides in the controller.

(c) *RabbitMQ*: Distributed operating system modules are not tied together. This type of OS uses message queue system in order to communicate between the modules. Similarly, OpenStack private cloud uses RabbitMQ to communicate between nodes.

4.2 Hadoop on OpenStack

Hadoop cluster is built on OpenStack cloud. Hadoop cluster consists of master and slave nodes. In order to provide Hadoop as a service, Hadoop cluster has to be configured. In order to configure Hadoop cluster, virtual machines are used which are provided by OpenStack cloud. OpenStack cloud provides KVM hypervisor to launch virtual machines as shown in Fig. 6.

Hadoop Web site user interface is designed by using Django. Users have option to select size of the Hadoop cluster, for example, 2, 3, 4, 5, and RAM size of each node. When the user selects Hadoop cluster size and RAM size of the each node, nodes must be launched. In order to launch virtual machines, openstacksdk's API is used. Configuration of hadoop cluster on launched virtual machines is done using automation by scripts. For example, Hadoop cluster of size 3 involves one virtual machine to be a master and other two virtual machines to be slaves. Configuration of master and slaves is done by using scripts. To access Hadoop service, Master IP address and Private Key are provided to users. Users must use master virtual machine to submit their jobs for execution by using IP address and Private Key.

Table 1 OpenStack cloud configuration

Name	RAM (GB)	Secondary memory (GB)
Controller	3	50
Neutron	2	50
Compute 1	4	50
Compute 2	6	50

Table 2 Hadoop cluster configuration

Name	RAM (GB)	VCPU	Secondary memory (GB)
Master	2	1	50
Slave 1	1	1	50
Slave 2	1	1	50

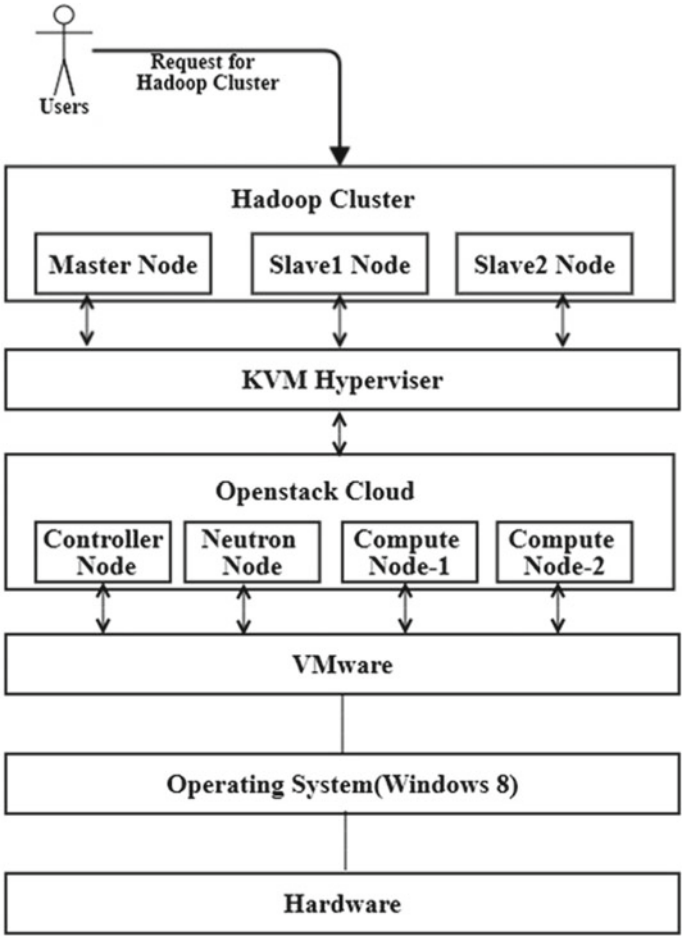

Fig. 6 Hadoop on OpenStack

5 Results and Discussion

OpenStack cloud consists of four virtual machines, i.e., 1 controller, 1 neutron, and 2 compute nodes (Table 1).

Table 3 Performance comparison of Hadoop cluster

Sl. no.	Type of Hadoop cluster	Size of the data (MB)	Execution time
1	On physical system	100	4 min 30 s
2	On OpenStack cloud system	100	5 min 20 s

Hadoop cluster consists of 1 master and 2 slave nodes. Hadoop cluster configuration such as RAM, VCPU, and secondary memory remains the same for the physical system and cloud system (Table 2).

After successful configuration of Hadoop cluster on physical system and openstack cloud system, word count program counts the number occurrences of each word in a file using MapReduce technique. It has been executed to test the performance of Hadoop cluster on physical and cloud system (Table 3).

Hadoop cluster on physical system is faster and efficient than openstack cloud system since processing is done without resource sharing on physical system as compared to openstack cloud system. Hadoop cluster on physical system is as powerful and fault tolerant compared to the openstack cloud system, but scalability is not easy as compared to openstack cloud system, because in openstack cloud system user can easily add virtual machines as per their requirement since the virtual machines are virtualized in cloud, and all the virtual machines reside in the same project, and in same network but whereas in physical system, it is not easy to add physical machines, as there is a network dependency among all nodes.

6 Conclusion

The proposed work focuses on providing Hadoop service to the user using openstack private cloud system by using openstacksdk's API and script. Performance analysis is carried out by considering Hadoop cluster on physical system and Hadoop cluster on OpenStack cloud system reveals Hadoop cluster on openstack cloud system is easily scalable since creation of virtual machines on cloud system is easy and another advantage is fault tolerant because if one node goes down or fails to perform as expected can be easily replaced by a new node. Even though the execution time for Hadoop cluster on cloud system is more compared with Hadoop cluster on physical system, data storage and flexibility are more in Hadoop cluster on cloud system.

7 Limitations and Future Scope

In this work, each node consists of 50 GB secondary memory and variable RAM size. Additional secondary memory cannot be allocated, if the user demands for more secondary memory. Since each OpenStack Virtual Machine consists of 50 GB default secondary memory.

OpenStack cloud consists of many components such as keystone, glance, nova, neutron, horizon, cinder, and heat. Cinder is the component, which provides volumes for virtual machines of OpenStack, i.e., secondary storage. This component can be added to OpenStack cloud to provide additional secondary storage for the users in future.

References

1. Varia J, Mathew S (2014) Overview of amazon web services. Amazon Web Services
2. Zheng Z, Zhu J (2013) Service-generated big data and big data-as-a-service: an overview. IEEE, pp 403–410
3. Palit HN, Dewi LP (2017) Exploratory research on developing hadoop-based data analytics. In: International conference on soft computing, intelligent system and information technology. IEEE, pp 160–166
4. Gaikwad C (2017) Providing storage as a service on cloud using OpenStack. In: International conference on innovations in information, embedded and communication systems. IEEE, pp 1–4
5. Hoefer CN, Karagiannis G (2010) Taxonomy of cloud computing services. IEEE, pp 1345–1350
6. Sheela PS (2017) Deploying an OpenStack cloud computing framework for university campus. In: International conference on computing, communication and automation. IEEE, pp 819–824