Comparison of Performance on Hadoop  Deployment Models

1Shweta V. Chaudhari, 2Revati M. Wahul
1,2Department of Computer Engineering, Modern Education Society College Of Engineering, Pune
Email: 1shwetavchaudhari@gmail.com, rmwahul@mescoepune.org

Abstract— Hadoop is an open-source Cloud computing environment that implements the MapReduce framework in Java. MapReduce makes it very easy to process and generate large data sets on the cloud. The main target of the cloud environment is power saving and resource utilization over unification while MapReduce works on large scale data analysis. This paper is for study of Hadoop energy efficiency by using physical and virtual clusters in different configurations.

Keywords— Cloud Computing, Hadoop, Virtualization

I. INTRODUCTION

Cloud computing and Hadoop MapReduce framework used to process the large data sets.

Hadoop is a framework, it will take into consideration the distributed processing of consequential data sets over clusters of PCs. It is intended to scale up from single servers to many machines, each offering nearby calculation and capacity.

As we can see in above diagram there are three main categories in the Hadoop development Client machine, Name node (Master node), Data node (Slave node).[2]

The Master nodes supervise the two key practical pieces that built Hadoop: storing large amount of data (HDFS), and running parallel calculations on all data (Map Reduce). The Name Node manages and organizes data storage capacity (HDFS), while the Job Tracker administers and arranges the parallel processing of information utilizing Map Reduce. Slave means both a Data Node and Task Tracker which is used to communicate with and accept the command from their master nodes. The Task Tracker work under the Data node and job tracker works under the Name Node.

The MapReduce function distributes file. Basically a large file is distributed in a block of identical size and they are split across the cluster for storage. [2]

In MapReduce implementation, there are three stages: Map, Shuffle, and Reduce. The map stage concern as map function to all input. It is utilized to handle the blocks in the input file that are keep into the PCs nearby capacity. Algorithm are done where the data are really put away. Since there is no any conditions in various mappers, all mappers do their work in parallel and they can work in parallel and separately to each other. In cluster if one computer fails to then result can be recomputed on another computer.

The real map function is called separately for each of these sets and makes a self-assertive expansive file of new key-value sets from it:

Map (key, value) -> List (key, value)

After Map function finishes its process it will pass the result to Shuffle function to arrange the resulting pair with their keys then pass it to Reducer as per their keys. The structure ensures all set with a similar key are appointed to a similar reducer. [2]

Now all pairs of key are gathered by reducer gather and creates a sorted list of the values. Input to the reduce function is key and the sorted list of values. To make a size of list very small, reduce function compact the list of values, it returns a single value as its output. The reduce function creates a list of key-value pairs, just like the map function:

Reduce (key, List(values)) -> List(key, value)

II. RELATED WORK

In this paper we will discuss previous work for performance and Energy Efficiency.

Performance Jeffery Shafer have analyze a few performance problem with the HDFS. Here we going to examine performance and energy utilization of Hadoop with its isolated data and compute services[3]. The author have experiment theVMM-Bypass I/O to enhance the execution time-basic on I/O operations, such systems are very helpful for the performance of Hadoop in virtualized enviornment [4].

Earlier task has demonstrated VMs are reasonable for executing data comprehensive for applications of Hadoop through utilization of sort or wordcount benchmarks[5]. Jian demonstrates that a appropriate
MapReduce execution can accomplish achievement near parallel databases over examinations performed on cloud service provider like Amazon EC2[6]. Earlier effort assessed Hadoop for logical applications or trades off of different equipment and file the framework configuration[7]. The previous efforts distinguished that doing experiments in virtualization become overhead with one VM per server [8].

**Energy Efficiency** Leverich demonstrates the Covering Subset (CS) data design and load balancing strategy.[9] An new model is called All-In Strategy (AIS) . AIS consider as a superior decision[8]. Earlier work demonstrates that DVFS can yield substantial energy savings in compute intensive Hadoop application [10]. Berkeley Energy Efficient MapReduce (BEEMR) proposes the preparing of collective jobs on little subspace of servers and moves the rest of the servers into a power compensating state[11]. In GreenHadoop research is done on opportunity of green energy and jobs energy requirements of MapReduce with in scheduling.[12]

### III. OVERVIEW

#### A. Hadoop Deployment Models

Different parameters that jobs need to have to be executed efficiently. These parameters are [13]:

- **Hadoop Parameters**: which is a set of predefined configuration parameters that are in Hadoop setting files.
- **Profile Statistics**: which are a set of user-defined properties of input data and functions like Map, Reduce, or Combine.
- **Profile Cost Factor**: which are I/O, CPU, and Network cost job execution parameters.

In this paper we will study Hadoop performance with collocated and separated data and compute services on physical and virtual clusters.

Let's start with deployment models, we will discuss two different deployment models first one is traditional model and second is alternate model.

In traditional Deployment model data and compute services are collocated and in alternate deployment model data and compute services are separated.[1]

#### B. Traditional Model

As we can see in fig2 Figure 2 shows traditional deployment model where data or compute services are collocated on any slave system. Slave system can be a server or VM.

Hadoop provide MapReduce Framework as well as Hadoop Distributed File System (HDFS). Job Tracker accept all request send by many client and their job send to data node to perform the job. If any condition Task Tracker get failure then this failure managed by Job Tracker. HDFS is used to store or manage I/O data. Name node is work as master node. It mange the metadata or many data nodes that holds the data blocks. The most important part of Hadoop is its capacity to evade the impacts of data locality. data locality empowers one to perform map where the data with in the Hadoop cluster located. In this way limiting the cost of input information developments to map tasks. TaskTracker and data node are collocated on each node.

### C. Alternate Model

In cloud computing machine are transient means they are workable only when any application need to execute. In alter deployment model data (data node) and compute service(Task Tracker) are separated. In this model services are run on dedicated node. While implementing this model the main foci will be on HDFS. Due to this separation in data and service Flexibility is provided by alternate deployment model.

### D. Effects of Hadoop Deployment Model

In previous work Hadoop is deployed on static cluster where all resources are controlled by Hadoop. Previously Hadoop is not designed for VMs, So Data and compute services are coexist. Traditional deployment models loss the elasticity property which is an important character of Cloud Computing.

In earlier one author start his work in elasticity in cloud environment with coexist in data and computing services Dynamically adding and removing the VMs are easy in cloud environment. Due to this it achieve the property automatic Scale Up/Scale Down the Hadoop cluster in runtime. VMs which are perform in HDFS become costly operation due to time is taken for adding and removing the VMs and it increase the overhead.

C. Advantages and Disadvantages of Hadoop Deployment Models
Advantages and Disadvantages of of Traditional Deployment Model

- Control over resource
- Easy to build cluster
- Significant performance degradation on VMs
- I/O heavy benchmarks perform poorly in virtualized environments
- Overhead compounded with multiple VMs per server
- Competing for resources

Advantages and Disadvantages of Alternate Deployment model

- Collocation consistently holds highest performance to power ratio
- Impact of isolating information and figure out services heavily depends on data compute ratio
- Adding more servers did not improvement

IV . CONCLUSION

In the paper discussion is started with performance of different deployment models of Hadoop with its advantages or disadvantages and performance is calculated by using execution time. Also, discussion with correlation with power consumption when Hadoop is running on physical or virtual cluster with data and computing services.

REFERENCE


understanding-hadoop-clusters-and-the-network/


