Asian Journal of Convergence in Technology Volume II Issue III
Issn No.:2350-1146, I.F-2.71

Green Computing using Hadoop : An Energy Efficient Technique

Sumedha M. Thakurdesai
Department of MCA, Finolex Academy of Management & Technology, Ratnagiri
sumedhathakurdesai@gmail.com

Abstract –
Cloud computing is an emerging model for distributed computing. Hadoop is the popular framework for handling large amount of data. It has two main components i.e. Hadoop Distributed File System and MapReduce method. The purpose of this paper is to implement this technology in the eco-friendly environment using green energy. This will reduce the electricity consumption and improve the performance of data processing.

In this way, we can save our environment or we will reduce the effect on environment by using Hadoop framework and Green renewable energy like wind and solar. This helps in protecting environment, reducing e-waste and increase in performance of data computing.

Index Terms - HDFS, Green energy, MapReduce, RAID, YARN, NameNode, Renewable enery

I. INTRODUCTION

In this age of big data, where the data volumes we need to work with on a day-to-day basis have outgrown the storage and processing capabilities of a single host. Big data brings with it two fundamental challenges i.e. (1) how to store and work with voluminous data sizes and (2) How to understand data and turn it into a competitive advantage [1].

Hadoop fills the gap in the market by effectively storing and providing computational capabilities for substantial amounts of data. It provides a way for distributed file system and parallel execution on distributed machines. The way Hadoop framework handles or manipulates or extract data we can say that it is eco friendly because it provides same CPU time to all job using JOB FAIR SCHEDULING algorithm to save power or work energy[2].

II. HADOOP

Hadoop is a platform that provides the distributed storage and computation capabilities. Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Cutting, who was working at Yahoo! at the time, named it after his son's toy elephant. It was originally developed to support distribution for the Nutch search engine project[7]

Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common and should be automatically handled by the framework.

The core of Apache Hadoop consists following primary components :
(a) Hadoop Distributed File System (HDFS) for data storage,
(b) Yet Another Resource Navigator (YARN) introduced in Hadoop 2. It is a general purpose scheduler and resource manager.
(c) A processing part called MapReduce. It is a batch-based computational engine.

Hadoop splits files into large blocks and distributes them across nodes in a cluster. To process data, Hadoop transfers packaged code for nodes to process in parallel based on the data that needs to be processed. This approach takes advantage of data locality— nodes manipulating the data they have access to— to allow the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking[2].

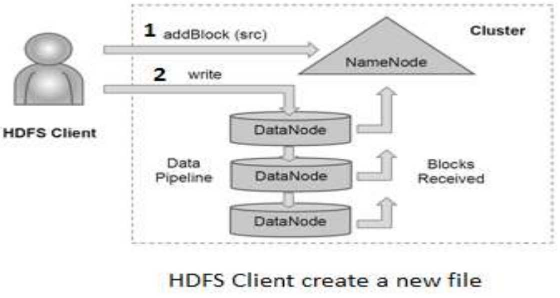

Figure 1: HDFS Communication Process

A) Components of Hadoop: [1]

1) Hadoop distributed file system : The Hadoop distributed file system (HDFS) is a distributed, scalable, and portable file-system written in Java for the Hadoop framework. Each datanode serves up blocks of data over the network using a block protocol specific to HDFS. The file system uses TCP/IP sockets for communication. Clients use remote procedure call (RPC) to communicate between each other.

HDFS stores large files (typically in the range of gigabytes to terabytes) across multiple machines. It is possible to replicate the data across multiple hosts, and hence theoretically does not require RAID (Redundant Array of Inexpensive Disks) storage on hosts With the default replication value, 3, data is stored on three nodes: two on the same rack, and one on a different rack. Data nodes can talk to each other to rebalance data, to move copies around, and to keep the replication of data high. HDFS added the high-availability capabilities, as announced for release 2.0 in May 2012, introducing the main metadata server (the NameNode).

Because the namenode is the single point for storage and management of metadata, it can become a bottleneck for supporting a huge number of files, especially a large number of small files. HDFS Federation, a new addition, aims to tackle this problem to a certain extent by allowing multiple namespaces served by separate namenodes.

An advantage of using HDFS is data awareness between the job tracker and task tracker. The job tracker schedules map or reduce jobs to task trackers with an awareness of the data location. E.g. if node A contains data (x,y,z) and node B contains data (a,b,c), the job tracker schedules node B to perform map or reduce tasks on (a,b,c) and node A would be scheduled to perform map or reduce tasks on (x,y,z). This reduces the amount of traffic that goes over the network and prevents unnecessary data transfer. When Hadoop is used with other file systems, this advantage is not always available. This can have a significant impact on job-completion times, which has been demonstrated when running data-intensive jobs[3].

2) YARN : It is Hadoop's distributed resource scheduler. It is added in the Hadoop version 2. It is used to deal with the challenges with the Hadoop architecture such as :

>> Deployments of larger than 4000 nodes will lead to the scalability issue.

>> To run the execution models such as machine learning algorithms that require the iterative computations.

Its main role is to schedule and manage resources in a Hadoop cluster. The components of YARN-ResourceManager, NodeManager, application client and container- can be described with the following figure :

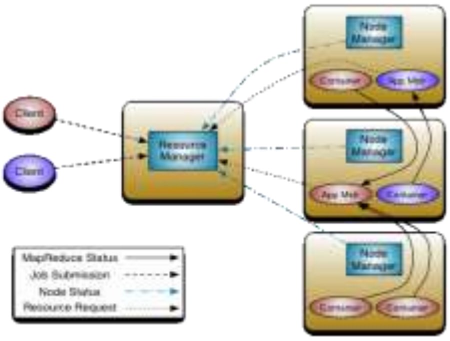

Figure 2: YARN Architecture

3) MapReduce: It is a batch-based, distributed framework to implement the parallel processing of the data. MapReduce decomposes work submitted by the client into small parallelized maps and reduce tasks as shown in the figure.

The role of the programmers is to define map and reduce functions where the map function outputs key/value pairs, which are used by reduce functions to deliver the final output.

The power of MapReduce occurs between the map output and the reduce input in the shuffle and sort phases as shown in figure :

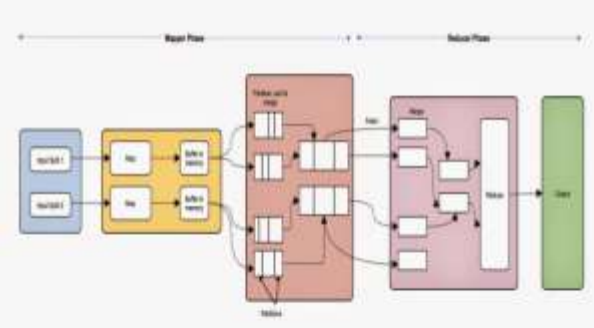

Figure 2 : Map & Reduce Functions

B) Some Applications of Hadoop:[4]

1) Hadoop technology in Cancer Treatments and Genomics: The objective of utilizing Hadoop as a part of Healthcare is to gather and dissect information that can do everything from survey general wellbeing patterns in a district of a huge number of individuals to pinpoint treatment choices for one tumour patient. There are around 3 billion base matches that constitute the human DNA and it is vital for such a lot of information to be composed in a compelling way in the event that we need to battle growth. The malignancy has not been cured yet is a direct result of the way that disease transforms in various examples and responds in various routes taking into account the hereditary cosmetics of a person. Consequently patients should be given customized treatment in view of the kind of tumor the individual patient's hereditary qualities make up. Utilizing Hadoop innovation will offer incredible backing for parallelization and help in mapping the 3 billion DNA base sets utilizing MapReduce programs.

PRESIDENT OBAMA'S PRECISION MEDICINE INITIATIVE WOULD
HELP DEVELOP BETTER TREATMENTS FOR DISEASES LIKE CANCER BY:
• Accelerating the design and testing of effective treatments tailored to individual patients
• Expanding genetically based clinical cancer trials
• Establishing a national "cancer knowledge network" to guide treatment decisions

Figure 3: Hadoop in cancer treatment

2) Hadoop technology in Monitoring Patient Vitals:

There are several hospitals across the world that use Hadoop to help the hospital staff work efficiently with Big

Data. Without Hadoop, most patient care systems could not even imagine working with unstructured data for analysis.

Children's Healthcare of Atlanta used a sensor beside the bed that helps them continuously track patient signs such as blood pressure, heartbeat and the respiratory rate. These sensors produce large chunks of data, which using legacy systems cannot be stored for more than 3 days for analysis. The main motive of Children's Healthcare of Atlanta was to store and analyze the vital signs. If there is any change in pattern, then the hospital wanted an alert to be generated to a team of doctors and assistants. All this was successfully achieved using Hadoop ecosystem components - Hive, Flume, Sqoop, Spark, and Impala[6].

Figure 4: Patient monitoring and Big Data

3) Hadoop technology in the Hospital Network: A Cleveland Clinic spinoff organization known as Explorys is making utilization of Big Data in medicinal services to give the best clinical backing, diminish the expense of consideration estimation and deal with the number of inhabitants in at-danger patients. Explorys has apparently constructed the biggest database in the medicinal services industry with over a hundred billion information using Hadoop. It offers clinicians some assistance with determining the deviations among patients and the impacts medications have on their wellbeing. These bits of knowledge help the restorative professionals and social insurance suppliers discover the best treatment gets ready for an arrangement of patient.

4) Hadoop technology in Healthcare Intelligence:

In the Healthcare Insurance Business the data and the outcomes are always dynamic and changing. Using Hadoop technology in Healthcare Intelligence applications helps hospitals, payers and healthcare agencies increase their competitive advantages by devising smart business solutions. For example, to find the age in a region below which there are no victims of certain disease and the compute the cost of policy, it requires to process huge data sets with the information like medicines, diseases, symptoms etc. In this case Hadoop's Pig, Hive and the MapReduce is the best solution.

5) Hadoop technology in Fraud Prevention and Detection

At least 10% of the Healthcare insurance payments are attributed to fraudulent claims. Big Data Analytics helps healthcare insurance companies find different ways to identify and prevent fraud at an early stage. Using Hadoop technology, insurance companies have been successful in developing predictive models to identify fraud customers by making use of real-time and historical data of medical claims, weather data, wages, voice recordings, demographics etc. The increasing demand for using Hadoop technology in Healthcare will eliminate the concept of "one size fits all" kind of medicines and treatments in the healthcare industry.

III. Green Energy

Natures colour will be green if and only if environment of the earth will be clean. Due to industrialization and urbanization Earth's environment is degrading day by day and second by second. Green energy is used to minimize the negative impact on the environment. Traditional energy sources, most notably fossil fuels, produce greenhouse gases that are believed to be the primary cause of an effect known as global warming or climate change.

A) Different Energy Resources:

Renewable energy is generally defined as energy that comes from resources which are naturally replenished on a human timescale, such as sunlight, wind, rain, tides, waves, and geothermal heat. Renewable energy replaces conventional fuels in four distinct areas: electricity generation, air and water heating/cooling, motor fuels, and rural (off-grid) energy services. Worldwide investments in renewable technologies amounted to more than US$214 billion in 2013, with countries like China and the United States heavily investing in wind, hydro, solar and biofuels.

B) Green Computing :

Green computing is the environmentally responsible and eco-friendly use of computers and their resources. Green computing includes the implementation of best practices, such as energy efficiency central processing units (CPUs), peripherals and servers. In addition green technology aims to reduce resource consumption and improve the disposal of electronic waste (e-waste)

C) Hadoop as a Source for Green Computing :

Map Reduce Technique can be used as Eco Friendly Techniques. Map Reduce is not using general job scheduling algorithm like FCFS, SJF, PRIORITY SHEDULING, ROUND ROBIN SCHEDULING etc. to handle user's job. Fair scheduling is a method of assigning resources to jobs such that all jobs get, on average, an equal share of resources over time. When there is a single job running, that job uses the entire cluster. When other jobs are submitted, tasks slots that free up are assigned to the new jobs, so that each job gets roughly the same amount of CPU time. While default Hadoop scheduler forms a queue of jobs, this lets short jobs finish in reasonable time while not starving long jobs. Fair sharing can also work with job priorities - the priorities are used as weights to determine the fraction of total compute time that each job gets. So, we can also customize fair scheduling algorithm on the basis of our requirements.[6]

Figure 5: Solar Energy

In this way we can say that every user's jobs get same amount of time whether user is special one or general which saves users time and energy as well as electric power used. That is why it is Eco friendly in nature. There is several green energy techniques like. Solar energy, Wind energy, Bio gas energy etc. But Solar energy and Wind energy are mostly used in industries as green energy techniques because they do not disturbs environment and don't produces waste material like thermal power plant. As we know that Solar/Wind techniques produces DIRECT CURRENT. So we have to use DC/AC inverter to convert DC to ALTERNATIVE CURRENT. The data center can generate its own SOLAR ENERGY using solar power computer system. We want to save energy by generating because of following reasons:

1) Because energy loses due to power generation and transmission is about 40%.
2) The ability to survive grid outage in developing countries.
3) Due to lower cost of establishment[2].

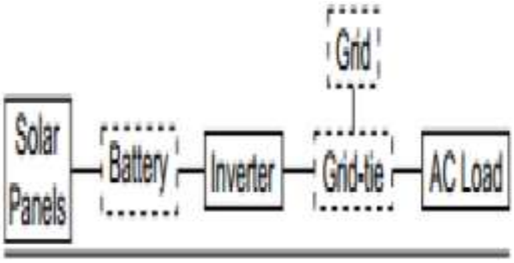

Figure 6: Solar Power Computer System

IV. Hadoop as a Service

- On-Demand Elastic Cluster: A Hadoop group in the cloud depends upon information handling necessities. Hubs are consequently added to or expelled from groups relying upon information size to enhance execution.
- Integrated Big Data Software: Hadoop as a Service incorporates full coordination with the Hadoop biological system, including MapReduce, Hive, Pig, Sqoop, Spark and resto.
- Simplified Cluster Management: Qubole Data Service offers a completely oversaw Hadoop-based group, wiping out the requirement for additional time and assets committed to overseeing hubs, setting up bunches and scaling framework.
- Lower Costs: Hadoop in the Cloud requires no forthright interest in on location equipment or IT sup

A) Implementation of Hadoop using Green Energy :@Rutgers University, New Jersey

Figure 7: Rutgers University, New Jersey

Researchers at the Polytechnic University of Catalonia and the Rutgers University have recently proposed building a programming framework that can effectively manage a datacenter's workload in a solar powered system. Dubbed the "Green Hadoop", the system will also be using the grid, but only as a backup source of power. The Rutgers team has developed their own Green Hadoop managed data centre using a small solar powered unit that they built and named "Parasol". The Parasol includes a container, a solar photovoltaic system, and a battery for storing energy. The datacenter can be switched to different sources of energy (complete grid, off grid, and hybrid) using three switches.

The technology is rapidly being employed by big name players, with Apple recently building a 20 MW solar panel farm in Maiden, NC in order to power their data centers. Facebook and eBay have their own smaller solar farms as well [6].

CONCLUSION

It has been observed that the fusion of the green energy and Hadoop will provide the environment friendly results and performance. Therefore, the use of this combination will be beneficial and eco-friendly for long term results.

REFERENCES

[1] Hadoop in Practice (second edition) by Alex Holmes , DreamTech Press
[2] *IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 4, Ver. VI (Jul – Aug. 2014), PP 52-56* Eco-Friendly Hadoop Shiv Kumar1, Shrawan kr.shrama2,Hitesh kumar swarnkar3
[3] https://en.wikipedia.org/wiki/Apache_Hadoop
[4] https://www.dezyre.com/article/5-healthcare-applications-of-hadoop-and-big-data/85
[5] https://www.qubole.com/hadoop-as-a-service/?nabe=6725258257104896:1&utm_referrer=https%3A%2F%2Fwww.google.com

[6] http://cloudtimes.org/
[7] http://hadoop.apache.org/
[8] http://hadoop.apache.org/docs/r1.0.4/fair_scheduler.html