Indian Journal of Science and Technology, Vol 10(10), DOI: 10.17485/ijst/2017/v10i10/107943, March 2017
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Integrated Hadoop Cloud Framework (IHCF)

Kumar Abhishek, Manish Kumar Verma, Kumar Shivam, Vinit Kumar and Adarsh Mohan

Department of CSE, NIT Patna, Ashok Rajpath, Mahendru, Patna - 800005, Bihar, India; kumar.abhishek@nitp.ac.in, manish123977@gmail.com, vinitkumar697@gmail.com, kshivam747@gmail.com, adarsh.mohan2011@gmail.com

Abstract

Objective: The paper proposes the design of an intermediate Hadoop framework which will not only make Hadoop user friendly but also it is open source and free to use. **Methods:** The framework has been called Integrated Hadoop Cloud Framework (IHCF) and it also supports various Hadoop based frameworks like Hive, Pig as cloud services. It can be accessed from outside the Hadoop cluster too. Various experiment results has been included which show the efficient working of IHCF. **Findings:** The IHCF contains modules like Setup, Client and Cloud which are working in sync with one another. The setup module controls automated creation of cluster and client module provides users access to the cluster. The cloud module handles Hadoop based frameworks and ensures that user/client can use frameworks as cloud services. **Improvement:** The IHCF can be customized further to ensure optimized use of clusters and prevent over/under utilization of resources in cluster.

Keywords: Big Data, Cloud, Cluster, Hadoop, Hive, Pig

1. Introduction

Big data is a relative term defining huge amount of datasets. Basically, it is a collection of datasets that cannot be processed by traditional large database systems. In today's world big data has not retained itself just to data rather than has expanded itself to a subject. It involves various tools and techniques along with various frameworks. Big data contains or involves all the data produced by different kinds of website and different kinds of devices and applications. They basically include:-

- Data by social media - The data generated by social media like Facebook, Yahoo, Twitter which includes the posts of billions of people across the world generates zeta bytes of data daily.
- Data by Stock exchange - The stock exchange which basically includes data of all the buy and sell decisions of shares of different companies and their profit and losses made by their customers.
- Black Box Data - Black box is a major device in helicopter, jets, airplanes etc. which records the voices of the pilots and crew and stores the information of aircraft performance, earphones and microphones.
- Power Grid Data - Power grid data contains information about power consumed by each

Figure 1. Different Kinds of Big Data[9].

*Author for correspondence

node with respect to power generated by base station.

- Transport Data - Transport data has all the transport information like model number, distance covered and availability of vehicle.
- Search Engine Data - Search engine data has all the data of different web crawlers, different search engines from different databases.

Figure 1 shows Different Kinds of Big Data[9]. The major challenges linked with big data are namely (i) Volume (ii) Variety (iii) Velocity (iv) Veracity (v) Validity (vi) Value (vii) Variability (viii) Venue (ix) Vocabulary and (x) Vagueness[1,2].

1.1 Big Data Solutions

1.1.1 Traditional Methods

According to the traditional method approach data is stored in relational databases like SQL, SQLite, MySQL server etc., and is accessed using the application or the interface. It suffers with limitations. While deploying this approach, we cannot store huge volumes of data. Also the data that is to be stored must be in a structured format. As Non-structured data cannot be stored in these databases that's the reason that data analysis takes a lot of time on relational data as every records are visited while looking for the right data.

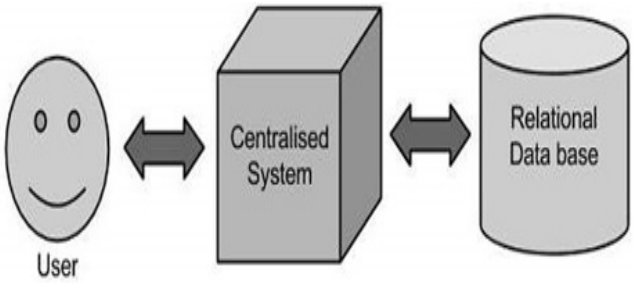

Figure 2. Traditional Approach[10].

1.1.2 Hadoop

Figure 2 shows the Traditional Approach[10]. Hadoop is an open source software framework used for computing huge datasets and distributed storage over large datasets present in clusters deployed by commodity hardware. Hadoop is based on Google's Map-Reduce programming model[15] including the file system model which was designed for the nutch search engine[16] project. Hadoop basically consists of two key parts - HDFS & Map Reduce. All the setup can be done on a single machine, but real deployment and the power of Hadoop lies within clusters of machines, as it can be expanded horizontally from a single node to multiple nodes in the cluster.

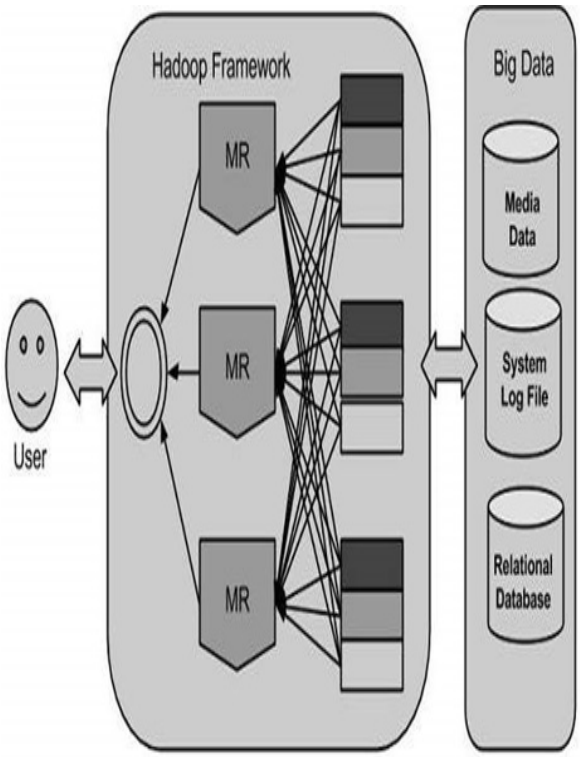

Figure 3. Hadoop Framework[10].

1.2 Need of Hadoop

Figure 3 shows Hadoop Framework[10]. The Apache Hadoop was developed as a solution to big data. As with the advancement in the technology. The data generated is increasing day-by-day. As a matter of fact 90 percent of the world's current data has been generated in the past few years only and it is still growing at exponential rate. The data produced till 2009 was 5 billion gigabytes and in 2011 this was the data produced in just two days and in 2013 it was the data produced every 10 minutes. And all this data cannot be neglected as they contain some useful information.

Hadoop also solves the vertical scaling problem[17] of the machines as the machines can now be scaled vertically and as many nodes can be added to the machine as per the need of the computation, speed, efficiency and storage. Figure 4 shows the importance of Hadoop[8].

Figure 4. Importance of Hadoop[8].

1.3 Hadoop Architecture

Figure 5 shows the Hadoop architecture[11]. It consists of the following four modules which are mentioned below:-

Hadoop Common: The Java libraries and various utilities which are needed by Hadoop and its modules are present in Hadoop common module. All these libraries consist of file system which is an Operating System level abstraction provided by Hadoop and it also contains all the necessary Java files and scripts which are needed to start Hadoop Cluster.

Hadoop YARN: Hadoop YARN is a framework which has been developed for job scheduling as well as cluster resource management. YARN refers to Yet Another Resource Navigator.

Hadoop Distributed File System (HDFS): Hadoop file system is a distributed file system which provides high-throughput access as well as high computation to application data.

Hadoop MapReduce: Hadoop MapReduce is a YARN-based programming paradigm which has been developed for parallel processing of large data sets.

1.4 Cloud

Cloud is resource which is not present on the client system itself but is present somewhere else and can be accessed as and when required as per the permissions granted by the server. As it is not present in the client side itself thus it is also said to be a virtual. The various services offered by cloud are namely Infrastructure as a service, Software as a service, and Platform as a service and Storage as a service. In infrastructure as a service complete infrastructure for running any kind of services is provided to the client whereas in platform as a service the particular platform required by the client is provided to them to work upon by the server. In storage as a service the requirement for a particular amount of storage is fulfilled whereas in software as a service the software required by the client is provided by the client.

According to NIST four types of cloud exist which are namely (i) **Private Clouds** – Maintained and organized by a company or organization for its work and is accessible only to the company, (ii) **Public Clouds** – Public cloud is accessible to as the name suggests to all the public, (iii) **Community Clouds** - This are made keeping in mind a common subsets of people with common needs and requirements and is accessible to them, and (iv) **Hybrid Clouds** - It is a hybrid or combination of all three i.e. public, private and community cloud.

Figure 6 illustrates the advantages of the cloud computing, some important of them are as follows-

- Reliability - Cloud is considered as a reliable source as data is backed up at many places by the server and the service provider also provides a 99% service level agreement.
- Cost Effective - Getting services through cloud is very much cost effective as compared to owning that resource if the need of the resource is not for much longer period.
- Manageability - Managing the resources of cloud are not the headache of the client. So manageability even being a big issue is not a thing to worry when it comes to cloud.
- Strategic Edge - In the competitive environment when an IT firm or company wants to be ahead of other company in terms of resources, cost and other factors, cloud give them a leading edge over other companies.

This paper aim to integrate cloud and Hadoop framework, so that it would become simple and friendly for the users. The first experience remains easy and nice without any complexity which are happening on the background

and without thinking about the system's configuration and capacity.

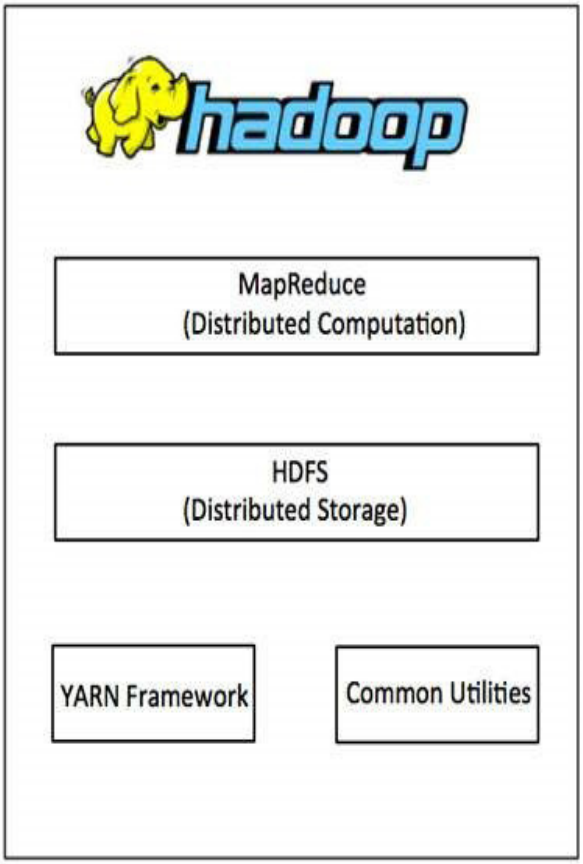

Figure 5. Hadoop Architecture[11].

This integrated framework will be light weight and generic in the type of task to be performed on cluster. Using this integration management of tasks of client will be easier without the need of any third party software.

It will increase efficiency and performance as users don't need to mug up large Hadoop commands to work on the cluster and they don't require thinking about their system configuration or any kind of software availability.

In[3] proposed a method of balancing, portability and performance of Hadoop distributed file system in 2010. The authors developed an approach to gain higher performance by efficiently managing the load, utilizing resources, decreasing overheads and minimizing the delay.

In[4] proposed the idea of map reduce as a simplified way of data processing in distributed environment for data processing on large clusters in the year 2008. Mapreduce is a programming paradigm and it is developed for processing data in distributed environment.

In[5] developed a high-performance distributed main memory based transaction processing system in the year 2008. The authors have implemented main memory transaction across distributed environments.

The paper[6] proposed an approach in the year 2009 about the new big data analysis practices and the proceedings of the VLDB Endowment. They analyzed about the new practices of how to make big data analysis more efficient.

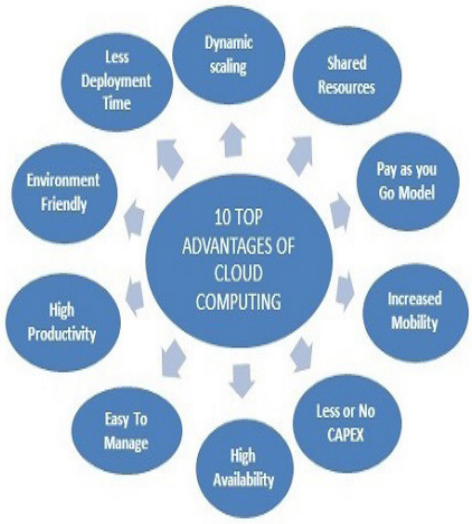

Figure 6. Advantages of Cloud[12].

The paper[7] proposed a way of how Apache Hadoop goes real time at Facebook. They gave real time analysis of Hadoop on Facebook over the real time data accessed by them.

After analyzing the above works, we conclude that Hadoop is becoming faster and efficient but no research has been carried to make Hadoop user friendly and reach the hands of common man. Hadoop has been confined to big corporate labs. Through Integrated Hadoop Cloud Framework (IHCF) we propose the design of a user friendly framework which will not only manage Hadoop clusters efficiently but is also open source and free. IHCF can be used to establish clusters of any size as required by user.

2. Proposed Framework

The Integrated Hadoop Cloud Framework (IHCF) follows N-Tier architecture as shown in Figure 7. The top

layer consists of presentation layer which mainly includes users and other software. The middleware is well connected to top presentation layer. The middleware has IHCF framework which is connected to both Hadoop based Frameworks and Cloud services.

Hadoop acts as base layer in middleware. Hadoop is directly connected to cluster of computers which form the resource layer for the framework. All operations are performed at this resource layer only.

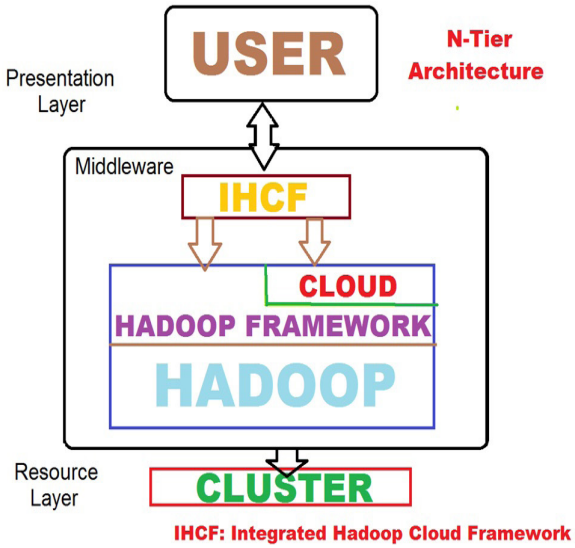

Figure 7. IHCF Architecture.

a. Additional Hadoop Based Frameworks in IHCF

i. Apache Hive

Apache Hive is a framework applied over Hadoop to execute ad-hoc queries on Hadoop cluster. Hive supports HQL (Hive Query Language) which is very much similar to SQL is used to process structured data present in data warehouses. It makes querying and analyzing of big data quite easy. Figure 8 shows Hive architecture.

ii. Apache Pig

Pig uses a high level procedural language called Pig Latin to process data in cluster. Pig Latin is very easy to write as its commands are very concise. It supports operations like join, group, filter, sort etc. Just like Hive Pig also executes on top of Hadoop. The pig engine converts Pig Latin codes to map reduce programs and execute them on Hadoop cluster. A user who is not familiar with Hadoop map reduce codes can work on big data clusters using Pig. Figure 9 shows the Pig components[14].

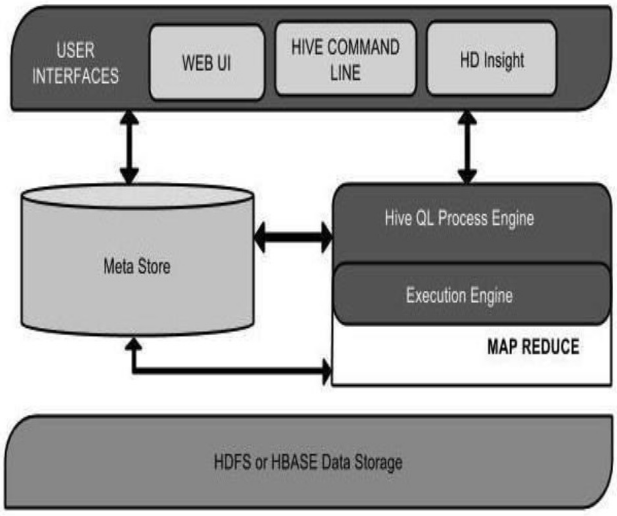

Figure 8. Hive Architecture[13].

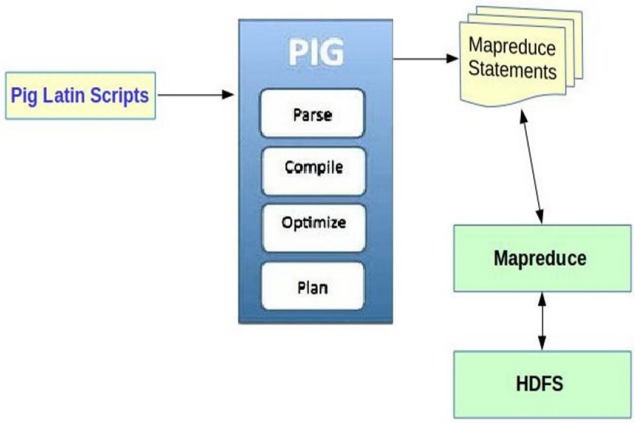

Figure 9. Pig Components[14].

The Apache Hive and Pig frameworks have been included in IHCF as cloud services where a client can use these frameworks even from a system which is not an actual part of Hadoop cluster. No installation of Hadoop or Pig is required on client system as these services are accessed from cloud server which is connected to Hadoop cluster.

3. Experimental Results

The Integrated Hadoop Cloud Framework (IHCF) has been implemented practically using shell scripts for framework design. VMware has been used to create a multi node cluster and deploy the IHCF on the cluster. Figures of various modules of IHCF have been included in this section.

a. Setup Module

The setup module has following tasks:-

- Automatic feasible Node Scanning in network:- The IHCF searches for nodes which can be added to cluster and it is shown in Figure 10.
- Selecting slaves and master as node's physical hardware configuration:- The nodes are assigned as master or slave depending on the resource configuration available on the node and it is shown in Figure 11.
- Auto installation of Hadoop over nodes:- The ICHF automatically installs Hadoop and other required components on the nodes of the cluster and it is shown in Figure 12.

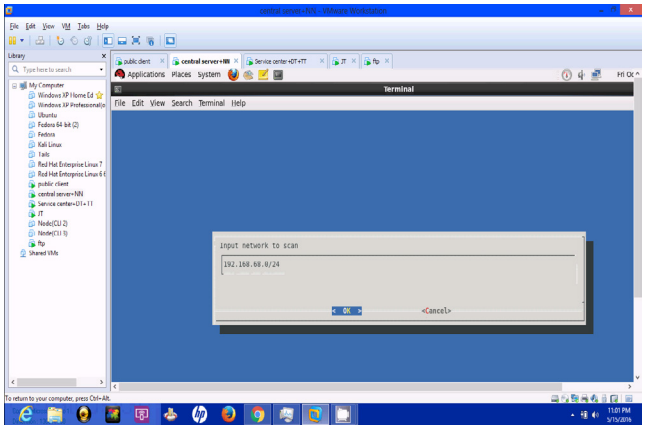

Figure 10. Automatic feasible Node Scanning in network.

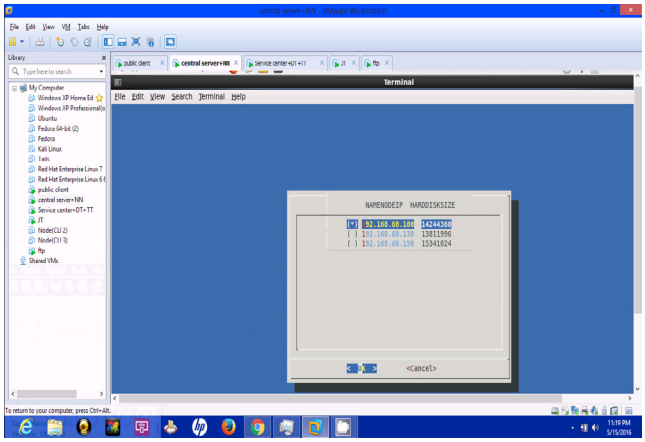

Figure 11. Selecting slaves and master as node's physical hardware configuration.

b. Client Module

The client module contains:-

- Login:- The clients login to IHCF using the username and password (as depicted in Figure 13).
- Cluster Status Check:- The client can check the status of resources inside the cluster (as depicted in Figure 14).
- Cluster File Upload:- The client can upload datasets to cluster via IHCF (as depicted in Figure 15).
- Cluster File Status Check: The status of uploaded datasets can be monitored by clients which is depicted in Figure 16.

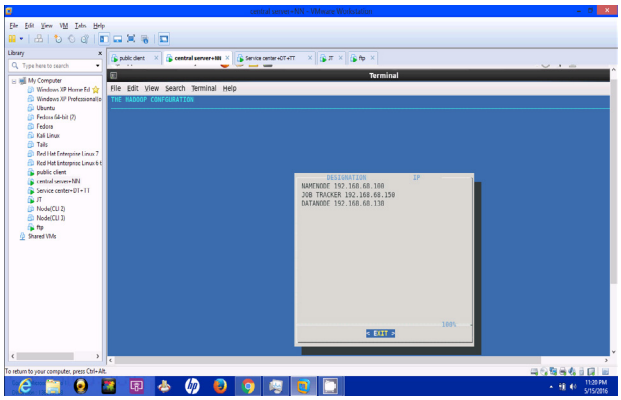

Figure 12. Auto installation of Hadoop over nodes.

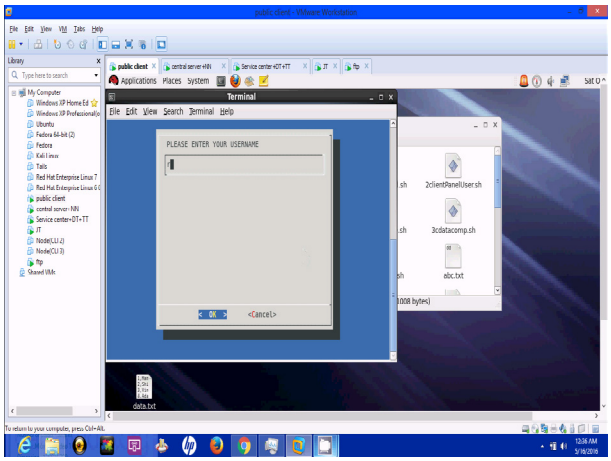

Figure 13. Client Login.

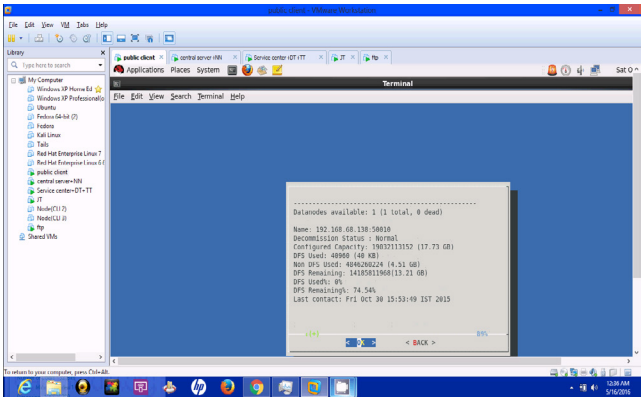

Figure 14. Cluster Status Check.

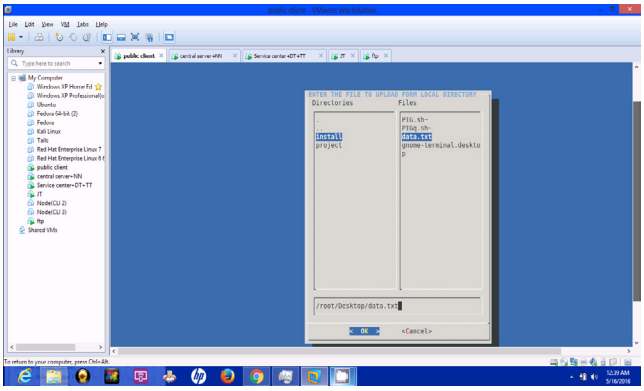

Figure 15. Cluster File Upload.

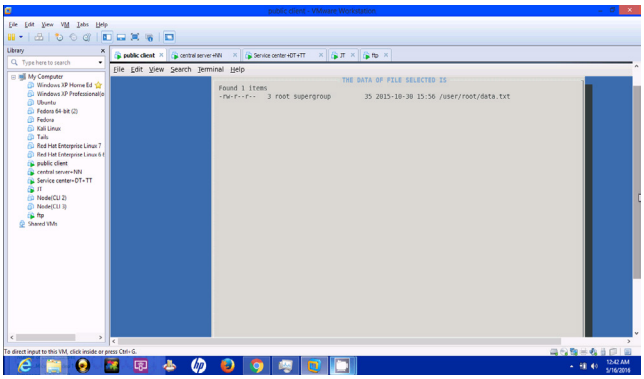

Figure 16. Cluster File Status Check.

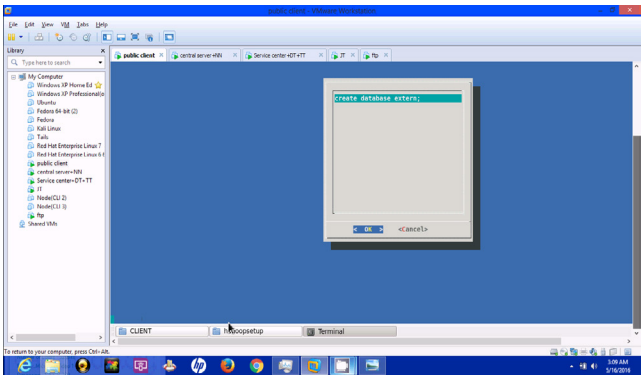

Figure 17. Hive Command.

c. Cloud Module

The cloud module contains various cloud based services:

- Hive Framework: Hive framework access has been provided as part of cloud service and it is shown in Figure 17.
- Pig Framework: Similar to Hive, Pig framework is also part of cloud service (as shown in Figure 18.
- Custom MapReduce: The IHCF give facility of custom Map Reduce so that the clients can write their own Map Reduce codes as per requirement. It is shown in Figure 19. Figure 17 shows the hive command and figure 18 shows the pig commands.

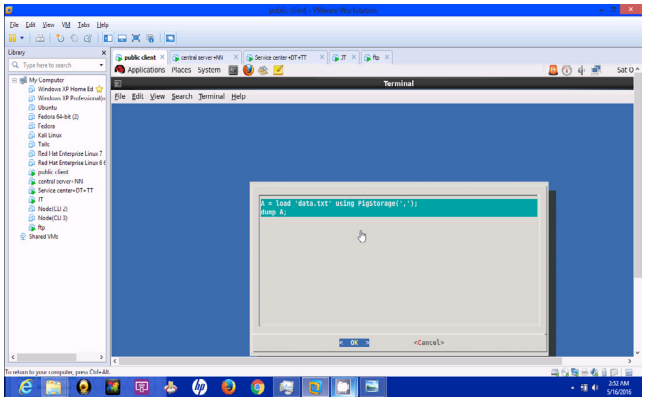

Figure 18. Pig Commands.

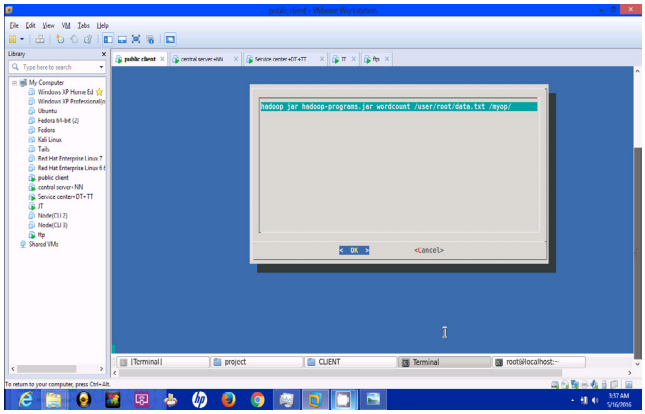

Figure 19. Custom Map Reduce.

4. Conclusion

Big Data is the outcome of evolving technologies and Hadoop is a possible solution to this issue. Using Hadoop and other frameworks it can effectively store large data and process them at very high rate and have much higher accuracy. Since Hadoop is open source, it is being widely used in various companies and enterprises to solve their big data problems. Hadoop has been a very efficient solution for many companies generating and analyzing peta bytes of data. It has solved various issues in industry linked to big data management and distributed systems. It is the future of data analytics.

IHCF will help to manage Hadoop more efficiently and effectively. Using this framework even a common man can use Hadoop.

5. Future Work

The paper can be further extended or improved by including the following features. For instance the client can access the cluster with some restrictions on resource uses. The client should not be given complete access to whole cluster as it might lead to resource misuse or mismanagement. Access to cluster must be synchronized in such a way that optimized use of resources can be done and there should not be any overutilization or underutilization of any resources.

Currently various task using map reduce are used manually by clients at the time of requirement. There should be a provision for task scheduling on cluster. It will benefit the client a lot as he/she need not to execute the map reduce program manually every time. This task execution can be automated using task scheduling.

6. References

1. Laney D. 3D data management: Controlling data volume, velocity and variety. META Group Research Note. 2001; p. 70.
2. Kirk Borne. Top 10 Big Data Challenges – A Serious Look at 10 Big Data V's. Available from: https://www.mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs.
3. Shafer J, Rixner S, Cox AL. The Hadoop distributed file system: Balancing portability and performance. 2010 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS). 2010; p. 122-33. Available from: Crossref. PMid:19563422.
4. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Communications of the ACM. 2008 Jan; 51(1):107-13. Available from: Crossref.
5. Kallman R, Kimura H, Natkins J, Pavlo A, Rasin A, Zdonik S, Jones EP, Madden S, Stone-Braker M, Zhang Y, Hugg J. H-store: a high-performance, distributed main memory transaction processing system. Proceedings of the VLDB Endowment. 2008; 1(2):1496-99. Available from: Crossref.
6. Cohen J, Dolan B, Dunlap M, Hellerstein JM, Welton C. MAD skills: new analysis practices for big data. Proceedings of the VLDB Endowment. 2009; 2(2):1481-92. Available from: Crossref.
7. Borthakur D, Gray J, Sarma JS, Muthukkaruppan K, Spiegelberg N, Kuang H, Ranganathan K, Molkov D, Menon A, Rash S, Schmidt R. Apache Hadoop goes real-time at Facebook. Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. 2011 June; p. 1071-80. Available from: Crossref.
8. What is Hadoop? Date Accessed: 12/12/2016: Available from: http://www.sas.com/en_us/insights/big-data/hadoop.html.
9. Hadoop-Big Data Overview. Date Accessed: 12/12/2016: Available from: http://www.tutorialspoint.com/hadoop/hadoop_big_data_overview.htm.
10. Hadoop - Big Data Solutions. Date Accessed: 12/12/2016: Available from: http://www.tutorialspoint.com/hadoop/hadoop_big_data_solutions.htm.
11. Hadoop - Introduction to Hadoop. Date Accessed: 12/12/2016: Available from: http://www.tutorialspoint.com/hadoop/hadoop_intro-duction.htm.
12. Top Ten Benefits of Cloud Computing Security Training. Date Accessed: 12/12/2016: Available from: http://www.simplilearn.com/cloud-computing-security-training-benefits-rar412-article.
13. Hive - Introduction. Date Accessed: 12/12/2016: Available from: http://www.tutorialspoint.com/hive/hive_introduction.htm.
14. Apache Pig Overview - Hadoop Online Tutorials. Date Accessed: 12/12/2016: Available from: http://hadooptutorial.info/apache-pig-overview.
15. Lammel R. Google's MapReduce programming model - Revisited. Science of Computer Programming. 2008 Jan; 70(1):1-30. Available from: Crossref.
16. Moreira JE, Michael MM, Da Silva D, Shiloach D, Dube P, Zhang L. Scalability of the Nutch search engine. Proceedings of the 21st annual international conference on Supercomputing. 2007; p. 3-12. Available from: Crossref.
17. Scalability. Date Accessed: 12/12/2016: Available from: https://en.wikipedia.org/wiki/Scalability#Horizontal_and_vertical_scaling.