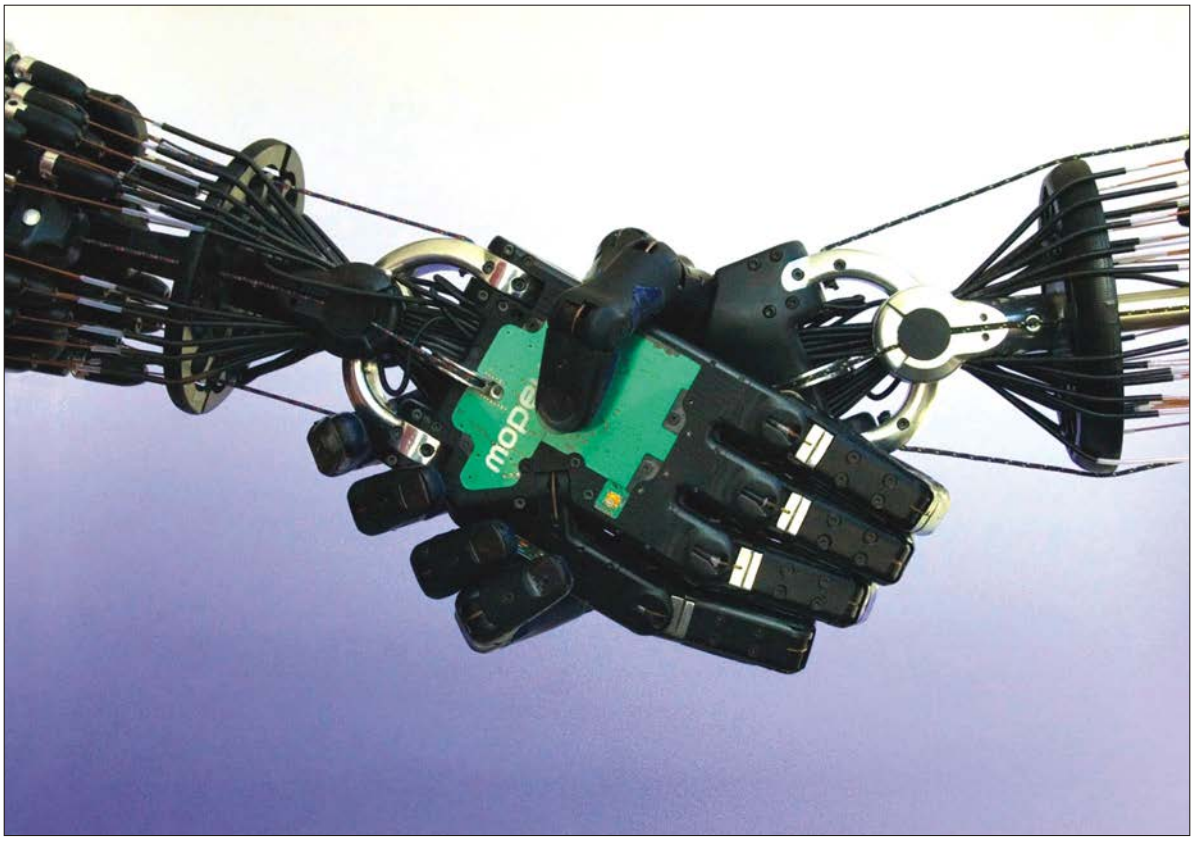

Journal of Artificial Intelligence

ISSN 1994-5450

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Artificial Intelligence 6 (1): 82-88, 2013
ISSN 1994-5450 / DOI: 10.3923/jai.2013.82.88

An Automation Tool for Single-node and Multi-node Hadoop Cluster

Vaibhav N. Keskar, Amit A. Kathade, Gopal S. Sarda and Amit D. Joshi
Department of Computer Engineering and IT, College of Engineering Pune, Pune-411005, India

Corresponding Author: Vaibhav N. Keskar, Sharada Apartment, Ward No-6, Near Kamgar Hospital, Shrirampur, Dist-Ahmednagar, 413709 Maharashtra, India Tel: +918087580935

ABSTRACT

Many organisations are required to deal with large data sets. To handle the large data sets these organisations use hadoop cluster. But they need to set up hadoop cluster with different number of nodes several times. This is very common case for many organisations that have just started with hadoop cluster. Setting up hadoop cluster is not a difficult job but requires large human efforts and time. As the number of nodes increases, human efforts and time required also increase. An automation tool which will automate the hadoop cluster set up procedure will be the possible solution to reduce human efforts and time requirement. This study talks about the tool which will automate the hadoop cluster set up procedure. No tool had proved a single operator to install single-node and multi-node cluster.

Key words: Single-node, multi-node, jobtracker, tasktracker, datanode, namenode, hadoop distributed file system

INTRODUCTION

Today's world is a world of large data, ranging from some petabytes to zetabytes. Many applications are required to deal with large data and the results should be obtained within a particular time limit. So, apache came with an appropriate framework to handle large data. This framework is Hadoop (Cloudera). It enables a number of applications to work with many computational independent machines and large data sets. Many leading companies like Yahoo, Facebook are using Hadoop.

Using a simple programming model Apache developed a framework for distributed processing of large data sets across the many computers or clusters of computers. This framewok is Apache hadoop software library. Hadoop (Cloudera) is designed to scale up from single machine to thousands of machines, each offering local computation and storage. Hadoop (Cloudera) itselfs detects and handles failures. Map/Reduce is a computational paradigm implemented by hadoop, where any job or application is divided into small fragments of work. These fragments may be executed on any node. Hadoop provides HDFS i.e., hadoop distributed file system which stores data on computer nodes. Both MapReduce and HDFS are designed in such a way that hadoop framework automatically handles the failure of nodes.

To set up a hadoop cluster requires some steps to follow which are listed in various hadoop (Cloudera) related documents (White, 2009). Single-node hadoop cluster is the basic form. In multi-node hadoop cluster, there is one master node and several slave nodes (Michael G. Noll). A small cluster has a single master node which consists of a jobtracker, tasktracker, namenode and datanode. Slave node consists of tasktracker and datanode.

Setting up the hadoop (Cloudera) cluster by manual method requires large human efforts and time. Many files need to be updated during installation. One possible method to reduce human efforts and time required for processing is to automate the process of installation. An automation tool will be one of the possible solutions. It uses the easiest approach to set up a multi-node hadoop cluster. Firstly, the tool sets up all the slave nodes as a single-node cluster, then the master node as a single-node cluster and finally combines all of them to make it a running multi-node hadoop cluster. The work gives an emphasis to design and develope a tool to install a single node and multinode cluster.

INSTALLATION STEPS

Many research studys and websites about hadoop cluster includes below mentioned installation steps:

Step 1: Add repository: Add repository in order to install jdk
Step 2: Install JDK and make it default for machine: Install jdk and set java home in ~/. bashrc file
Step 3: Add a dedicated user: Add a user account on ubuntu who will use hadoop file system
Step 4: D. SSH (OpenBSD) configuration: Create a connection between hadoop dedicated user and its local host
Step 5: E. Download hadoop: Download hadoop from apache mirror. Choose any version available
Step 6: F. Configure HDFS files: Update HDFS configuration files
Step 7: G. Format namenode: Format the HDFS via namenode

All the above mentioned steps are done manually in order to set up a single-node hadoop cluster. In order to set up a multi-node cluster, some HDFS files needs to be configured again and a connection needs to be established between every slave node and master node.

PROBLEM STATEMENT

To automate the single-node and multi-node hadoop cluster set up procedure with reduced human efforts and time required. Basic purpose is to set up a multi-node hadoop cluster with maximum possible number of nodes and minimum human efforts and time required. Single operator can set up the entire cluster.

PROPOSED SOLUTION

One solution to the above stated problem is to automate the process of installing hadoop cluster. Steps which are performed during installation can be automated into a tool which will only take the number of machines required by the user as input and set up a multi-node hadoop cluster automatically. It will automatically find the machines live on the network from its IP address. First, this tool will set up every node as a single-node hadoop cluster and then it will combine all those nodes to set up a running multi-node hadoop cluster. The process of setting up the single-node cluster on machines will run on every node in parallel. After installing the single-node hadoop cluster on every slave machine, a single-node cluster will be installed on the machine on which the tool is being run. This machine is considered as the master machine. After single-node cluster is ready on the master machine, the process of setting up multi-node cluster will begin. Finally, a multi-node cluster will be available for the user. This will save a lot of time and human efforts.

IMPLEMENTATION

Pictorial representation of software approach: Figure 1 shows the pictorial representation of the tool approach to set up a multi-node hadoop cluster.

Algorithm 1:

Ensure

1. OS-Ubuntu 9.04 and above
2. SSH (OpenBSD) enabled machines
3. Homogenous Environment machines

1. Start the tool
2. Select the hadoop version to be downloaded
3. Enter the number of slave nodes required
4. Identify own IP address
5. **while**(live IP found != number of slaves) **do**
6. **if** (ping next IP address) **then**
7. add to file1
8. live IP count++
9. **end if**
10. **end while**
11. Open file1
12. **while** (file1 !=EOF) **do**
13. copy all required files to a machine whose **IP** address is stored in file1
14. connect to a machine whose **IP** address is stored in file1 and start installation and continue
15. **end while**
16. start Installation on the master machine
17. create connection between master and all slaves
18. format namenode
19. start dfs
20. start mapred
21. STOP

Algorithm 2 (Leidner and Berosik, 2009):

1. add repository
2. Unzip the hadoop.tar.gz
3. copy hadoop to /usr/local
4. download JDK
5. add dedicated user
6. connect to localhost
7. update .bashrc of dedicated user
8. update HDFS configuration files
9. namenode format
10. start single node
11. stop single node
12. update HDFS configuration files for multinode set up
13. update /etc/hosts file

Download hadoop: Download any of the available version of hadoop. User can select any version from the graphical user interface of the tool and use for the installation porpose.

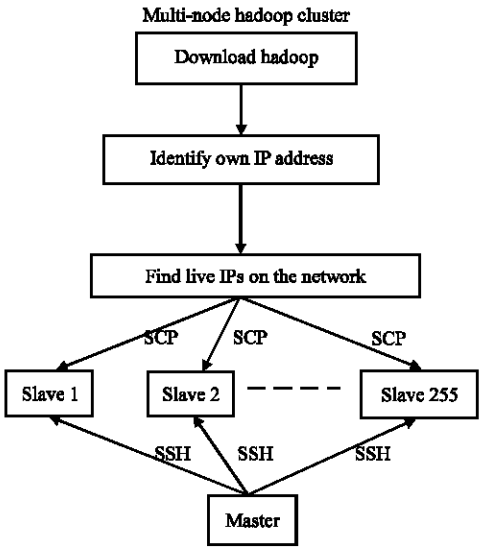

Fig. 1: Multi-node hadoop cluster set up-automation tool approach

Take user input about no of slaves: The only input required is the number of slave nodes that user wants for hadoop cluster.

Identify own IP address: Tool will identify IP address of the machine where the tool is running. This IP address is used to identify the other live machines on the network.

First come first serve (FCFS): As stated above, IP address of the machine where the tool is running is used to search live machines on network. The first machine encountered alive on the network is taken first and processing starts there. Subsequent machines are discovered similarly. IP addresses of all the live machines on the network are stored in a file.

Secure copy (SCP): Using the stored IP addresses one by one, tool transfers the files and shell scripts required during installation on machine having that IP address using SCP utility (Secure Copy) (Algorithm 1).

SSH: OpenBSD: When file transfer is complete, a new terminal window opens. This terminal window shows the installation process running on a perticular machine. A shell script (Algorithm 2) sets up a single-node hadoop cluster.

Add repository: Tool adds repository for installing sun-Java6-JDK using apt-add-repository command available in shell. Adding repository, tool enables user to download and install the JDK.

JDK: After addition of repository, tool will install sun-Java6-JDK automatically. After successful installation of jdk on the system, it sets Java home variable to sun-Java6-JDK.

Dedicated user: Tool adds a new user account on ubuntu for hadoop use. Only this user is able to use the hadoop distributed file system that is HDFS.

SSH: OpenBSD localhost: Tool creates a RSA public/private key pair for the dedicated user created in the previous step. It adds its own public key to its authorized keys file under ~/.ssh directory so that dedicated user can connect to local host without being prompt for password.

HDFS configuration files: These configuration files are conf/core-sites.xml, conf/hdfs-sites.xml and conf/mapred-sites.xml. These files are present in hadoop directory. Tool configures these files with the desired values. After the configuration is complete it formats the HDFS system via name node. After the formatting the name node, machine is ready to use as a single-node hadoop cluster.

Master: After all other machines are ready as a single-node hadoop cluster, tool starts the same process at the machine where tool is running. It sets a single-node cluster at this machine. Tool refers it as master node because this machine acts as a master node in multi-node hadoop cluster after the complete installation.

Configuration for multi-node: At this point of time all slaves and a master are individual single-node hadoop cluster. In this step tool configures all slaves and master node to be able to act as multi-node cluster. For this tool configures the previously stated hadoop configuration file along with conf/slave, conf/master and /etc/hosts file. The hadoop configuration files and /etc/hosts file have to be configured at every machine. conf/master and conf/slave are configured only at master node. Hosts file includes the IP addresses and corresponding names of all the slave nodes. And finally it formats this file system via name node.

Starting and testing cluster: Now tool starts multi-node hadoop cluster using the shell scripts provided in hadoop files. To test the cluster it uses jps command which shows what processes are working on that node.

RESULTS

Graph plotted in Fig. 2 gives clear idea about the time requirement of both manual and automatic procedure for multi-node hadoop cluster set up. This graph is based on the expected readings.

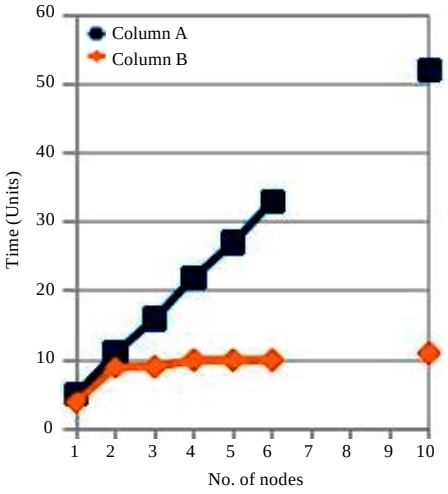

Fig. 2: Expected results-time requirement reduction due to automation

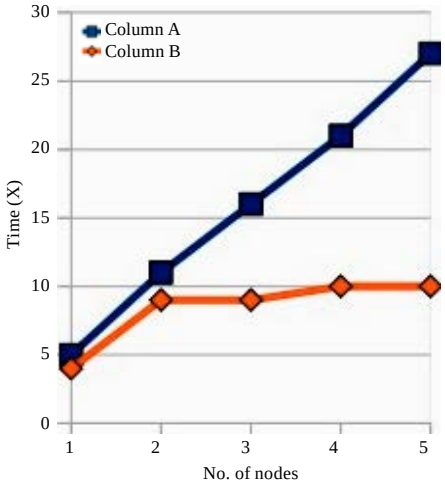

Fig. 3: Relative time variation between manual and automatic procedure of hadoop cluster set up for 5 nodes

All slave nodes installation procedure works parallel. So, actual time required to set up a multi-node cluster having more than two nodes is nearly equivalent to the time required to set up two-node hadoop cluster manually. This tool works in two phases. First Phase involves setting up all the slave nodes parallely, time required for this installation is nearly equal to time required to install single-node by manual method, since all the slave nodes are installed parallely. Second phase involves setting up master node, time required to install the master node is nearly equal to time required to install single-node by manual method. It means time required by this tool to install N nodes hadoop cluster is nearly equal to the time required to install 2 nodes hadoop cluster manually.

Suppose to set a single-node cluster manually time required is X units i.e., to set up N node hadoop cluster manually, time required is NX+T units where, T is a time variable which changes according to number of nodes such that T is directly proportional to N. To set the single-node cluster using tool, time required is X' units, such that X'<= X. But to set N node cluster using tool it requires just 2X'+T' time units where T' is a time variable and it changes with number of nodes such that T' is directly proportional to N. Time required to set N nodes hadoop cluster using tool is 2X'+T' time units because X' time units are required to set N-1 slave nodes and X' time units for a master node. Hence 2X'+T' is lesser than NX+T where N>2. Figure 3 shows the time variation for 5 nodes. These results may change according to internet speed but the relative time difference will be same.

CONCLUSION

This study gives clear idea about the tool and its effectiveness. This tool requires just one operator to set up the entire hadoop cluster. Human Efforts and time required are relatively much less than manual method. It will be an effective tool for many organisations just starting with hadoop cluster.

REFERENCES

Leidner, J.L. and G. Berosik, 2009. Building and installing a Hadoop/MapReduce cluster from commodity components: A case study. Technical Report. http://arxiv.org/ftp/arxiv/papers/0911/0911.5438.pdf

White, T., 2009. Hadoop: The Definitive Guide. 1st Edn., O'Reilly Media Inc., New York, USA., ISBN-10: 0596521979, Pages: 528.