HADOOP: How to Process the Big Data?

Ladislav Buřita[1], David Koblížek[2],

[1]University of Defence, Kounicova 65, 662 10 Brno, Czech Republic and
Tomas Bata University in Zlin, Mostní 5139, 760 01 Zlin, Czech Republic,
ladislav.burita@unob.cz
[2]Vojenský útvar 4854, Pardubice, Czech Republic and
University of Defence, Kounicova 65, 662 10 Brno, Czech Republic,
koblizekd@gmail.com

Abstract. The paper is concerned with scalable pre-processing of data using HADOOP that is a framework based on java for processing of large volumes of data, so called Big Data.. The first part is focused on explaining the main part of the HADOOP system, which includes distributed file system and MapReduce metod. In the second part is described the process of system installation and in the final part is explained the reason of the HADOOP experiment.

Keywords: HADOOP, Big Data, HDFS, MapReduce, Apache

1 Introduction

The paper is a particular result of the University of Defence (UoD) in Brno research project [1]. In the paper is explained the HADOOP structure with the distributed file system and the MapReduce method. Next is described process of installation of Apache HADOOP to Ubuntu linux server and in the final part is explained the reason of the experiment.

Apache HADOOP is a software platform based on java that is used to process very large volumes of data (Big Data) and their storage. HADOOP is prepared to run on a large number of low-cost machines that don´t have enough memory or disks. This solution is known like distributed system. Individual stations do not share anything and work independently. This allows arbitrarily add stations to the cluster, for examples if we need increase capacity or in case of hardware failure. The HADOOP architecture is different in the philosophy of data transfer. Client only sends MapReduce program for data, which saves the data transfer. [2]

2 The Apache HADOOP Architecture

The Apache HADOOP has two main parts: MapReduce and HDFS (HADOOP Distributed File System). MapReduce is a framework for processing nodes in a cluster. HDFS spans all the data nodes in a cluster for data storage. It links together

the file systems on many local nodes to make them into one big file system. HDFS assumes nodes will fail, so it achieves reliability by replicating data across multiple nodes [5][6].

2.1 The HADOOP Structure

The HADOOP is used for distributed file system and distributed computing architecture master/slave. The main instances in HADOOP architecture are NameNode and JobTracker [2].

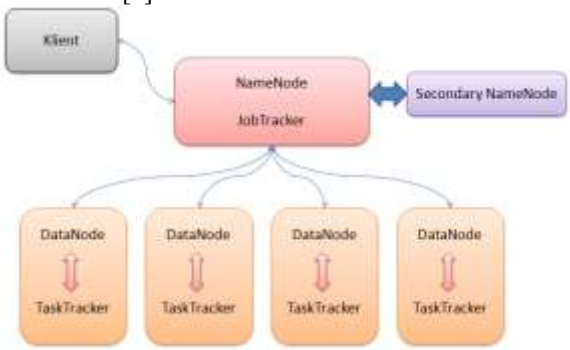

Fig. 1. The HADOOP structure (source [2])

NameNode serves as a repository for metadata of the shared file system, which include important information about the structure of files and directories of DataNodes. The server (on which runs NameNode instance) is also called HDFS master, because the slave station DataNode enters commands to perform basic input/output operations. The server, which runs NameNode, contains no data for processing and is not adapted to the computationally intensive tasks. It is the only place for HADOOP cluster.

Secondary Namenode is a backup instance for monitoring HDFS file system. Each HADOOP cluster contains one Secondary NameNode, which run on separate computer because of the possibility of restoration failure. Unlike NameNode, Secondary NameNode does not get any information about changes in the file system in real time. Secondary NameNode communicates with NameNode for creating backup metadata in the time interval, which can be set in a cluster configuration.

DataNode service provides storage for the shared file system. All instances of DataNode managed a block storage system for HDFS. There is only one instance in HDFS DataNode for each computer. DataNodes communicate with each other, there is a data replication to other nodes.

JobTracker provides a direct connection between the user application and HADOOP cluster. After sending MapReduce jobs to the cluster performs JobTracker completele management and planning. Specifies, which data will be processed, assign tasks to individual computers and also monitors their course. Int the event of failure or other problem JobTracker task automatically restarts and starts to another computer that is in order. There is only one instance of this server in the cluster and usually runs on the same computer with NameNode. JobTracker, also called MapReduce master, is

control element which oversees the execution of MapReduce jobs as a whole.

TaskTracker manages individual tasks at each station. Each TaskTracker, which runs on all substations in charge of running individual tasks allotted JobTracker. If a job is fails again JobTracker schedules the execution of the task to another functional TaskTracker. There is only one instance running on the computer.

2.2 The Distributed File System HDFS

HDFS is a distributed file system designed to run MapReduce jobs with large amounts of input data commonly available and inexpensive hardware.

In case of a request to create a file that does not automatically direct contact NameNode server, but to save the data to a local disk to a location for temporary data using HDFS client. An application that allows you to write is transparently redirected to this place until the stored data exceeds the size of one HDFS block. Subsequently, the client contacts the NameNode server, which inserts the file name in the file hierarchy system and assigns it to data block.

The NameNode sends the client identification to DataNode and target block of data at the end of the file. The client flushes the place to store temporary storage and server must inform the NameNode that closes the file. At this point NameNode marks the data as a permanent change. If the server NameNode report closure received or failed alone, would be lost file [2].

3 The HADOOP Installation on Ubuntu

In this section is described the basic process of installation and setup HADOOP on Ubuntu linux server. HADOOP can be configured in a single-node and multi-node cluster. Here we show single-node setup where all instances running on a single computer.

3.1 The Preparation Phase

1) Install the Sun Java 6 [3, 4]:

```
$ sudo add-apt-repository "deb
http://archive.ubuntu.com/ubuntu hardy main multiverse"
$ sudo add-apt-repository "deb
http://archive.ubuntu.com/ubuntu hardy-updates main
multiverse"
$ sudo add-apt-repository "deb
http://archive.canonical.com/ lucid partner"
$ sudo add-apt-repository "deb
http://ppa.launchpad.net/webupd8team/java/ubuntu
precise main"
$ sudo apt-get update
```

```
$ sudo apt-get install sun-java5-jdk sun-java6-jdk
oracle-java7-installer
```

2) Add a HADOOP system user:

```
$ sudo addgroup HADOOP
$ sudo adduser --ingroup HADOOP hduser
```

3) Connect ssh to localhost:

```
user@ubuntu:~$ su - hduser
hduser@ubuntu:~$ ssh-keygen -t rsa -P ""
hduser@ubuntu:~$ cat $HOME/.ssh/id_rsa.pub >>
$HOME/.ssh/authorized_keys
hduser@ubuntu:~$ ssh localhost
```

4) Set up hduser the root rights:

```
$ sudo adduser hduser sudo
$ /usr/sbin/visudo
root ALL= (ALL:ALL) ALL
hduser ALL= (ALL:ALL) ALL
```

5) Disable IPv6 in the file /etc/sysctl.conf:

```
net.ipv6.conf.all.disable_ipv6 = 1
net.ipv6.conf.default.disable_ipv6 = 1
net.ipv6.conf.lo.disable_ipv6 = 1
```

3.2 The Installation Phase

1) Download HADOOP from the Apache mirror and extract files:

```
$ cd /home/hduser
$ sudo tar xzf HADOOP-1.1.2.tar.gz
$ sudo mv HADOOP-1.1.2 HADOOP
$ sudo chown -R hduser:HADOOP HADOOP
```

2) Open `$HOME/.bashrc` and add the following line.

```
export HADOOP_HOME=/home/hduser/HADOOP
export JAVA_HOME=/usr/lib/jvm/java-6-sun
```

3) Edit /home/hduser/HADOOP/conf/HADOOP-env.sh, add following line:

```
export JAVA_HOME=/usr/lib/jvm/java-6-sun
```

4) Configurate HADOOP files:

- conf/core-site.xml:

```
<property>
    <name>HADOOP.tmp.dir</name>
    <value>/app/HADOOP/tmp</value>
</property>
<property>
    <name>fs.default.name</name>
    <value>hdfs://localhost:54310</value>
</property>
```

- conf/mapred-site.xml:

```
<property>
    <name>mapred.job.tracker</name>
    <value>localhost:54311</value>
</property>
```

- conf/hdfs-site.xml:

```
<property>
    <name>dfs.replication</name>
    <value>1</value>
</property>
```

This is the end of installation and configuration of HADOOP.

5) Format the HDFS file system via NameNode:

```
hduser@ubuntu:~$ /home/hduser/HADOOP/bin/HADOOP
namenode –format
```

After formatting is the HADOOP ready for use in single-mode.

6) Enter the command and check all instances of HADOOP:

```
"hduser@ubuntu:/home/hduser/HADOOP$ jps"
```

The HADOOP installation process is not a trivial task. It is necessary work step by step, after intensive literature study, and in each step to evaluate and correct the results. The literature sources are not enough clear and complete, it still lefts experimental work.

4 Application of HADOOP

HADOOP is designed for processing large amounts of data on large computer clusters that are assembled from commercially available hardware. This solution brings significant cost saving because there is no need to buy expensive computing resources. In the absence of total data space is possible simply just connect to the cluster next computers.

Main advantage is the distribution of HADOOP computing power on a large number of computers. HADOOP also saves network traffic because the data are not transmitted, only compute commands.

In the Army of the Czech Republic (ACR) are a lot of systems that monitors various parameters, for examples radiation, chemical and biological danger. With HADOOP would significantly improve the efficiency of the system. The collected data could simply be written to the HDFS cluster and analysis would be more accelerated. The same situation should be solved by sensor data processing.

The other option would be using HADOOP to create a user interface for storing and analyzing meteorological information. In the university environment could be used the described distributed environment for the storage and use of information resources.

The project of the HADDOP application (within [1]) is prepared in a simple frame: to gain an expertise in the new concept of Big Data processing. The first step "The

HADOOP understanding and installation" was successful finished. The next step is to design and develop application for demonstration of system features within the dissertation (author Koblížek) and for the educational purpose (author Buřita).

5 Conclusion

In the first part of the article is explained the structure of HADOOP. In the second part is described the installation process of HADOOP cluster in a single-node mode, where all instances running on the same computer.

In a real deployment it is necessary to set up HADOOP in multi-node mode, so that the individual computers running instance according to their position in the HADOOP structure. The installation of multi-node mode is the same as the single-node mode, but it is necessary to set all individual computers in the network environment (master / slave).

Some problems were appeared, when installed on the virtual Ubuntu server, but all were eventually solved, se the instruction set in the chapter 3. When setting up multi-node mode, no further problems occurred, so relatively quickly was build HADOOP cluster consists of 4 VMs.

The HADOOP is to be in the ACR easily and rapidly deployable due to the solved project and with help of number of commercial software solutions. The process of the HADOOP adoption will continue in the design of an application to demonstrate effective using of the system.

References

1. Project of the Institutional support to the Ministry of Defence, Czech Republic, to development of the research organization - University of Defence (UoD), Faculty of military technology, CIS department "Advanced technologies in communications and information systems", subproject: "Information and knowledge management in NEC environment". UoD: Brno, Czech Republic (2011-2015)
2. Marinič, M.: Scalable preprocessing of data using HADOOP tool. Bachelor's thesis. Brno, University of technology (2012)
3. Running HADOOP on Ubuntu linux, http://www.michael-noll.com/tutorials/running-HADOOP-on-ubuntu-linux-single-node-cluster/ (2013)
4. How to install HADOOP on Ubuntu on Single Node cluster, http://mohsin-junaid.blogspot.cz/2013/02/how-to-install-HADOOP-104-on-ubuntu.html (2013)
5. HADOOP, http://HADOOP.apache.org/ (2013)
6. What is HADOOP?, http://www-01.ibm.com/software/data/infosphere/HADOOP/ (2013)