An Approach to Setup Hadoop in Windows Environment

Hemant Hingave
PG Student, Dept. of Computer Science and Engg.
Yeshwantrao Chavan College of Engineering,
Nagpur (MS), India
hemant.hingave@gmail.com

Prof. Rasika Ingle
Dept. of Computer Technology
Yeshwantrao Chavan College of Engineering,
Nagpur (MS), India
meet.rasika@gmail.com

***Abstract*—**Hadoop is a open source framework for automatic parallelization of computing tasks in distributed environment. Unfortunately programming for Hadoop comprise of certain challenges. It is very difficult to debug and understand Hadoop programs. We can make it a little simple by using a simplified version of the Hadoop cluster that runs locally on the developer's machine. In this paper we describes how to set up such hadoop a cluster on a pc running Microsoft Windows. It also describes a way to integrate this cluster with a major Java development environment.

Keywords*-Hadoop, Marven, Cygwin, Namenode, Datanode*

*****

I. INTRODUCTION

Hadoop is a free, Open source Java-based programming framework which used to process of large data sets in a distributed computing environment or in cluster. Hadoop is part of the Apache project sponsored by the Apache Software Foundation[4]. Hadoop works on thousands of nodes involving thousands of terabytes of data. HDFS(Hadoop Distributed File System) is distributed file system facilitates rapid data transfer rates between nodes and allows the system to continue operating uninterrupted in case of a node failure by feature of fault tolerance[4]. Hadoop was inspired by Google's MapReduce[6].

Apache Hadoop 2.x release supports for running Hadoop framework on Microsoft Windows environment. But due to deficiency of some windows native components (like hadoop.dll, winutils.exe, etc) in bin distribution of Apache Hadoop 2.x release encounter ERROR util.Shell: Failed to locate the winutils binary in the hadoop binary path.[1]

Thus, we describe how to build bin native distribution from source codes, How to install, How to configure and run Hadoop in Windows Platform. Since hadoop is very complex environment we broke this approach into several smaller steps of configuration. Each step involves particular action will execute on setting some aspect of the system.

II. IMPLEMENTATION

Hadoop cluster Can be setup in the three supported modes. Local (Standalone) Mode, Pseudo-Distributed Mode and Fully-Distributed Mode. By default, In non-distributed mode hadoop work as a single Java process. This is useful for debugging perspective. In this approach, we setup hadoop in a pseudo-distributed mode where each Hadoop daemon runs in a separate Java process.

A. Prerequisites

- Microsoft Windows SDK v7.1.

The Windows SDK provides tools like headers, libraries, code samples and compilers a new help system that developers can use to create applications that run on Microsoft Windows.

- Cygwin.

Cygwin is an Open Source tools which provide functionality similar to a Linux distribution environment on Windows. The substantial POSIX API functionality provided by a DLL (cygwin1.dll).

- Maven 3.1.1.

Apache Maven is a software project comprehension and management tool. Supported the concept of a project object model (POM), Maven can manage a project's build, reporting and documentation from a central piece of information. Maven is a build automation tool used build Java projects in .jar.

- Protocol Buffers 2.5.0

Protocol Buffers are used for serializing structured data. They are provides interface for developing programs to communicate with each other over a wire or for storing data. The method that describes the structure of some data and a program that generates from that description source code in various programming languages for generating or parsing a stream of bytes that represents the structured data.

B. Set Environment Variables

Setup the Environment Variables as JAVA_HOME, M2_HOME and Platform. Variable name are case sensitive. The value for Platform variable will be either x64 or Win32 for building on a 64-bit or 32-bit os. JDK installation path should not contains any space for the

JAVA_HOME environment variable. After Edit Path Variable sited at installation directory to add bin directory of Cygwin (say *C:\cygwin64\bin*), bin directory of Maven (say *C:\maven\bin*) and installation path of Protocol Buffers (say *c:\protobuf*). The Fig 1 describe overall process.

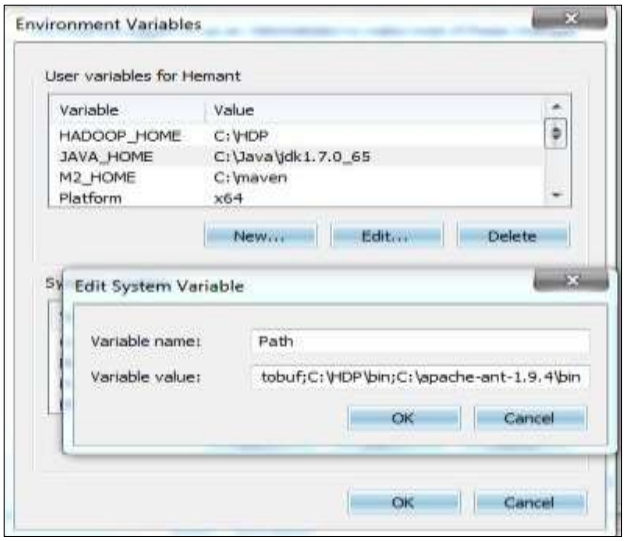

Fig 1. Setup Environment Variable

C. *Download hadoop and place it in the home directory.*

After that download hadoop-2.x-src.tar.gz and extract compress file to a folder having shorten path (say c:\hdfs) to avoid runtime problem due to maximum path length limitation in Windows.

D. *Create Windows binary tar distribution for native support*

The native support for Hadoop is provided by mvn package. To setup environment Select Start → All Programs → Microsoft Windows SDK v7.1 and open Windows SDK 7.1 Command Prompt. Change directory to Hadoop source code folder (say c:\hdfs). Execute mvn package with options -Pdist,native-win -DskipTests -Dtar to create Windows binary tar distribution. After process completion native distribution hadoop-2.x.tar.gz will be created inside C:\hdfs\hadoop-dist\target\hadoop-2.x directory.

E. *Install Hadoop*

Download the apache hadoop from source distribution or mirror. Extract hadoop-2.x.tar.gz to a folder (say c:\hadoop). Add Environment Variable HADOOP_HOME and edit Path Variable to add bin directory of HADOOP_HOME (say C:\hadoop\bin).

F. *Configure Hadoop*

following changes are mandatory to configure Hadoop.

- *File: C:\hadoop\etc\hadoop\core-site.xml-* It stores name and URL of the default file system. The uri's authority is used to determine the port, host etc. for a filesystem. We set its value as hdfs://localhost:9000
- *File: C:\hadoop\etc\hadoop\hdfs-site.xml-* ***dfs.replication:*** Specified Default block replication factor. The replications factor can be specified when the file is created and when replication is not specified the default value is used. ***dfs.namenode.name.dir*:** This specify directory where name node store the name table(fsimage) on the local filesystem DFS, ***dfs.datanode.data.dir***: This specify directory where data node in DFS on the local filesystem store its blocks.
- *File C:\hadoop\etc\hadoop\yarn-site.xml-* **yarn.nodemanager.aux-services**:
  This is auxiliary service name. Its Default value is omapreduce_shuffle **yarn.nodemanager.aux-services.mapreduce.shuffle.class**: This is auxiliary service class to use. The Default value is org.apache.hadoop.mapred.ShuffleHandler
  yarn.application.classpath: This specify classpath for YARN applications.
- *File: C:\hadoop\etc\hadoop\mapred-site.xml-* mapreduce.framework.name:
  The runtime framework for executing MapReduce jobs. Can be one of local, classic or yarn.

G. *Format the namenode*

Namenode should be formatted for the first time only on setup node.

H. *Start the cluster*

Start HDFS: The Namenode and Datanode can be started by executing command '\sbin>start-dfs'. Fig 3 Shows Namenode and Datanode

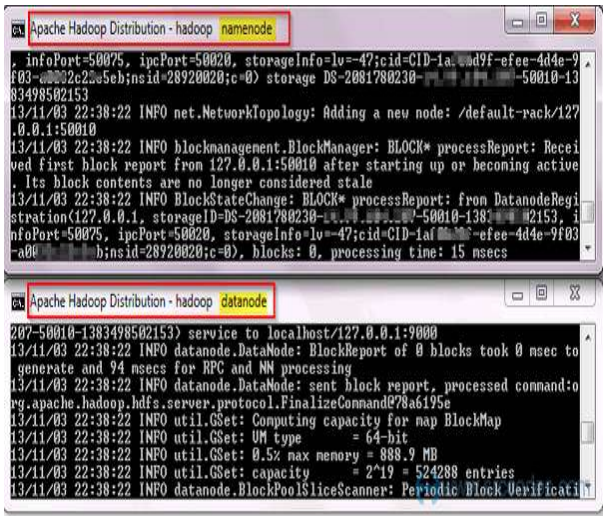

Fig 2 Namenode and Datanode

 ISSN: 2321-8169
1117 - 1119

Start MapReduce aka YARN: YARN split up the two major functionalities of the JobTracker, resource management and job scheduling/monitoring, into separate daemons as Resource Manager and Node Manager.

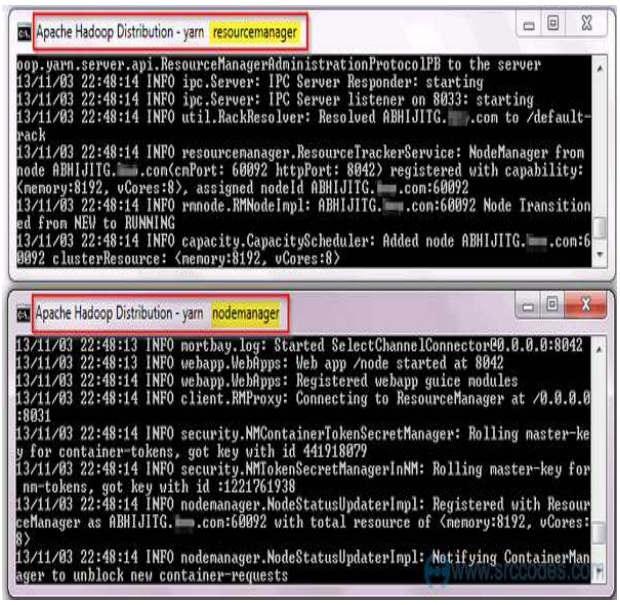

Fig 3 Resource Manager and Node Manager

Fig 3 Shows two separate Command Prompt windows one for Resource Manager and another for Node Manager.

I. Verify Installation

Finally we shows Resource Manager and Node Manager at http://localhost:8042 and Namenode at http://localhost:50070.

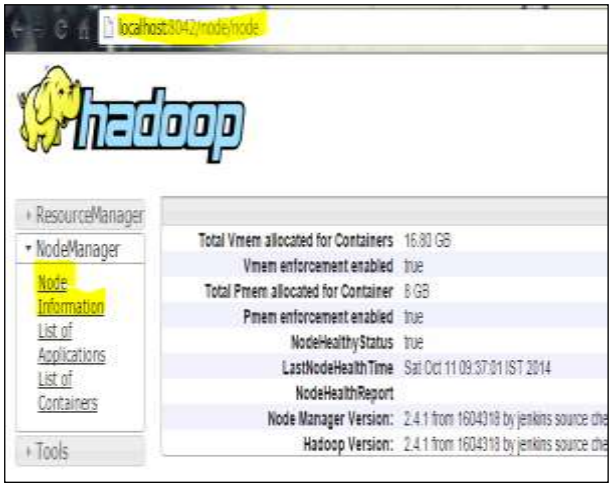

Fig 4. Namenode Information

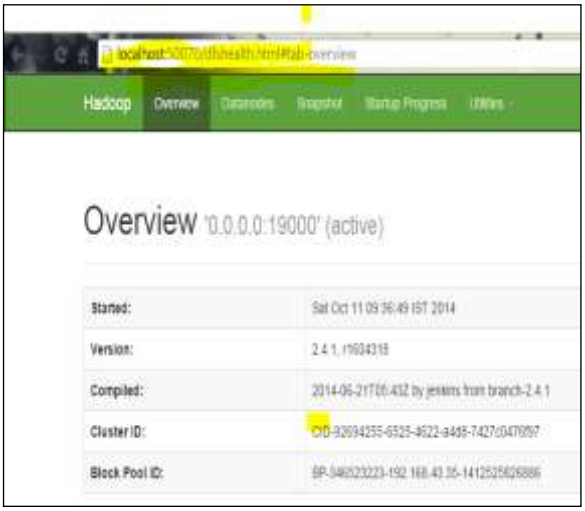

Fig 5 DfsHealth Status

III. CONCLUSION

This approach is simple to setup and run hadoop application efficiently. From the developer perspective, Hadoop is relatively straight forward. The hardest part of MapReduce programming is understanding how to translate algorithms and standard processing techniques into their MR equivalents. Windows environment provides flexibility to configure and debugging MapReduce program easily.

REFERENCES

[1] Abhijit Ghosh. "Build, Install, Configure and Run Apache Hadoop 2.2.0 in Microsoft Windows OS". Internet: http://www.srccodes.com/p/article/38/build-install-configure-run-apache-hadoop-2.2.0-microsoft-windows-os , Nov 3, 2013 [Oct. 03, 2014].

[2] Vlad Korolev, "Hadoop on Windows with Eclipse". Internet: http://v-lad.org/Tutorials/Hadoop/00%20-%20Intro.html , 2008 [Oct. 07, 2014].

[3] Arpit Agarwal, "Hadoop2OnWindows". Internet: http://v-lad.org/Tutorials/Hadoop/00%20-%20Intro.html , Aug 27 2014 [Oct. 08, 2014].

[4] EMC Academic Alliance, "DEEP DIVE INTO HADOOP". Internet: http://www.video28.com/video/jDOE8zx0Wsg/deep-dive-into-hadoop.html , [Oct. 08, 2014].

[5] O'Brien, et al., Tim. "Maven: The Complete Reference". Sonatype, 15 March 2013.

[6] Wikipedia , "Apache Hadoop". Internet: https://wiki.apache.org/hadoop/ , [Oct. 08, 2014].