IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Volume 6, PP 27-33

Deploy Hadoop For Processing Text Data To Run Map Reduce Application On A Single Site

Shrusti Talati[1], Dr. Mamta Meena[2], Shrutika Mallya[3]
[1,3]*(Student, Department of Computer Engineering, Atharva College of Engineering, Mumbai, India)*
[2, 3]*(Assistant Professor, Department of Computer Engineering, Atharva College of Engineering, Mumbai, India)*

Abstract: *The arrival of many ubiquitous devices, social networking and other sources of data has created a large amount of data with greater velocity and variety. Multiple organizations are applying big data analytics to challenges such as detection of fraud, analysis of risk, analysis of sentiments, analysis of equities, forecasting of weather, recommendations of various products and their classifications. So the big data is a collection of large datasets that cannot be processed using traditional computing techniques. Hadoop is a Java based open source platform that can process this data over thousands of distributed affordable commodity nodes and deliver predictive insights. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. In this paper we have installed and configured Hadoop on Windows platform. We have run an application in three wordcount languages , which are MapReduce.*
Keywords: *Hadoop, MapReduce, Big Data Analytics*

I. Introduction

We live in the age of data. People click pictures in their mobiles ,save videos, message friends, update their status on Facebook, comment on the web, redirect on ads, and so on. Machine Logs of Machine, RFID based readers, sensor network, GPS based vehicle trace, retail transactions etc contribute to the growing amount of data. Big data is a term for such sets of data that are very large or complex that traditional and old applications of processing the data are insufficient. Hadoop, is open source framework based on java that uses scale out approach for running distributed applications to exploit the power of commodity hardware rather than high end nodes. In this paper we have explained the process of configuring and deploying hadoop framework for processing text data and runninga MapReduce application on single site. The remaining paper is organized as follows: Section II describes the hadoop ecosystem. Section III explains the hadoop configuration process. Section IV discusses about development and execution of one MapReduce application on hadoop and Section V concludes the paper.

II. Configuring Hadoop

This section explains the process of installing and configuring Hadoop on windows.This section is mainly divided in two subsections such as building hadoop and configuring hadoop.

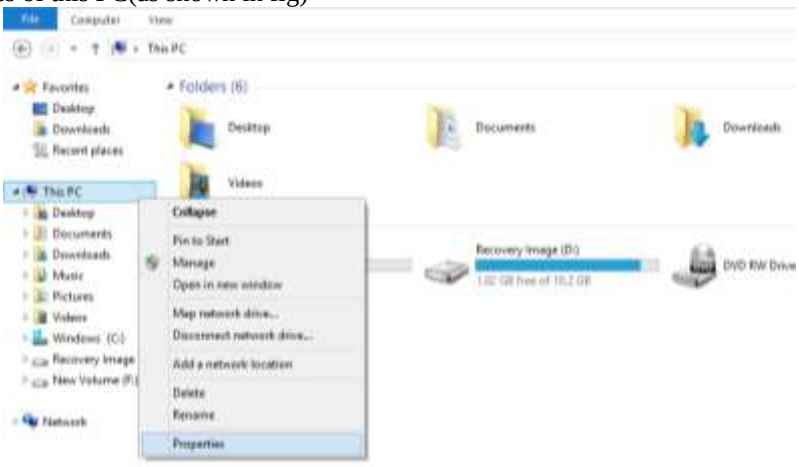
A. Build Hadoop

Download all the following & extract to c: drive.

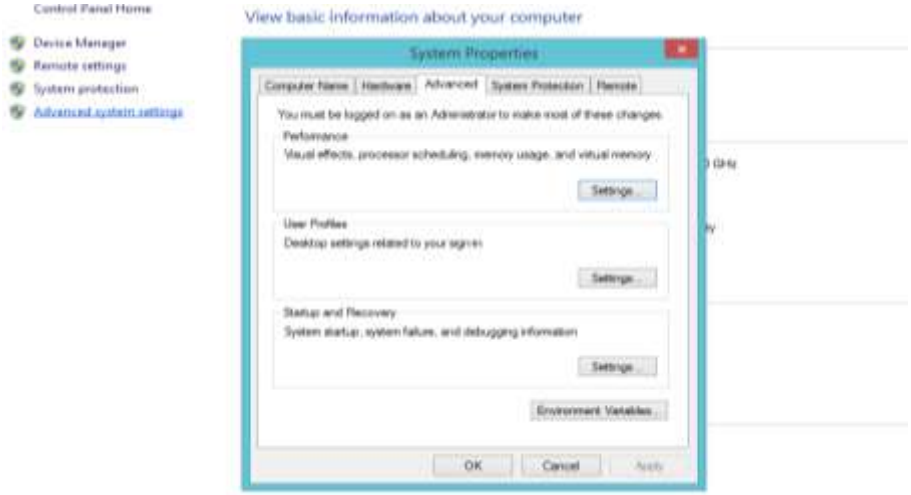
a. Download and installJDK 1.7
b. Download and install hadoop-2.7.1-src.tar.gz and change the folder name to hadoop-2.7.1 (short path)to avoid runtime problem due to maximum path length limitation in Windows
c. Download and install Maven 3.2.3
d. Downloadand install protobuf-master
e. Downloadand installpig-0.13.0
f. Downloadand installapache-hive-1.2.1-sr[2]

Now all the files are downloaded and now we'll set the path,
To set the PATH:

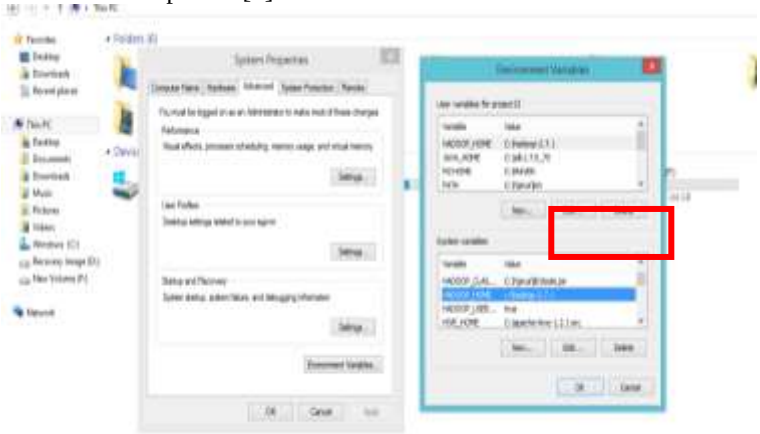
1. Go to properties of this PC(as shown in fig)



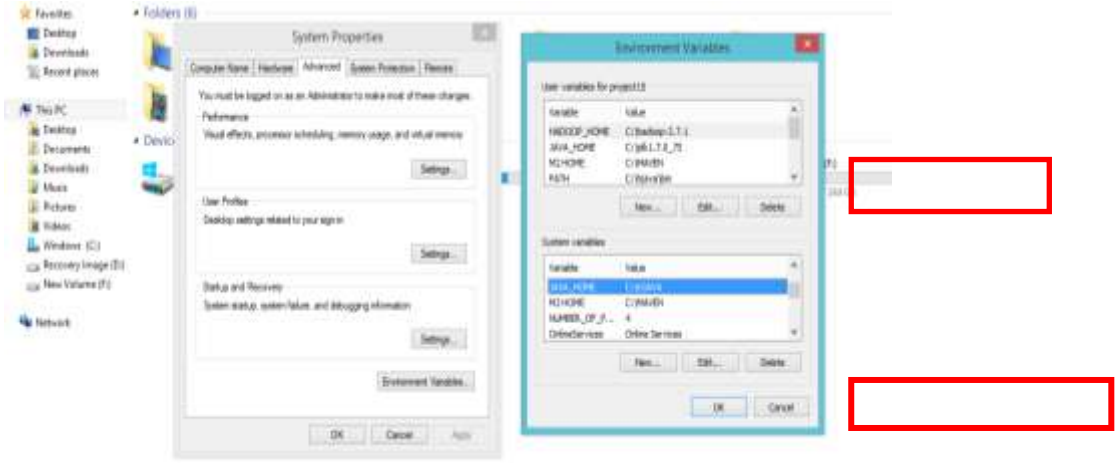
2. In properties select Environment Variables



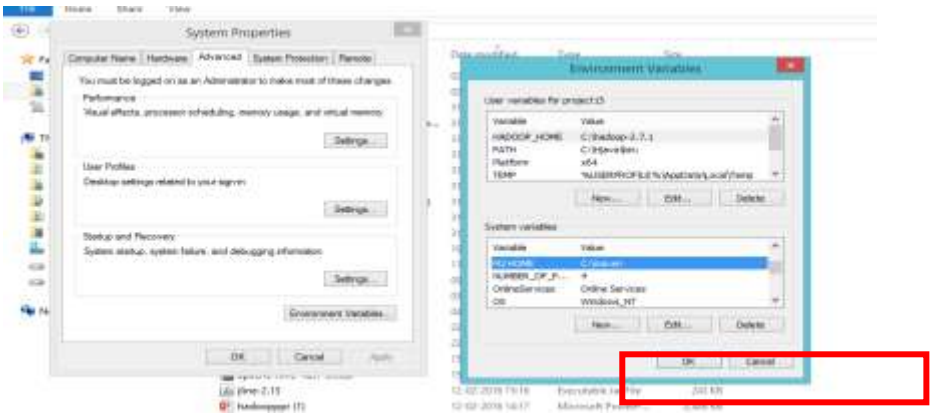
3. Now create new in system variables for HADOOP_HOME:
   In variable name: HADOOP_HOME
   Variable value: c:\hadoop-2.7.1[3]



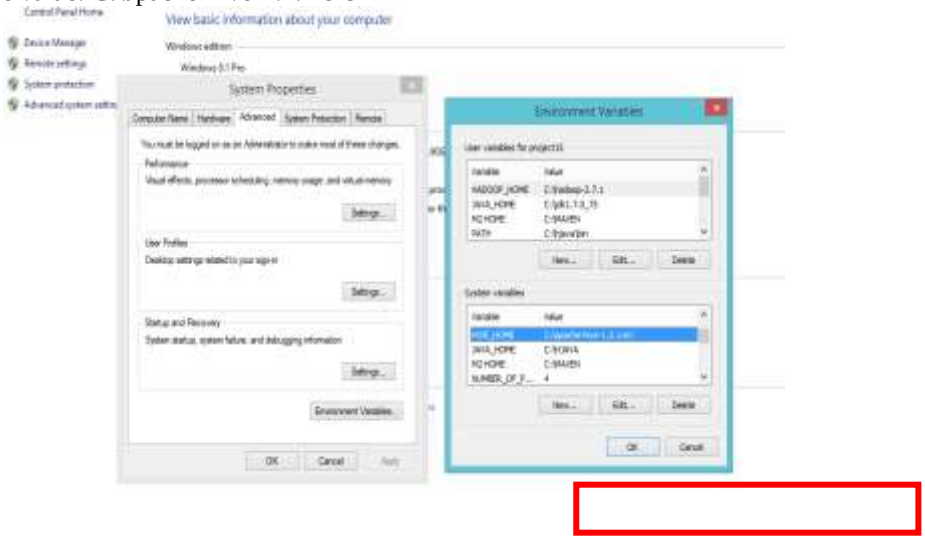
4. Now create new in system variables for *JAVA_HOME*:
   In variable name: *JAVA_HOME*
   Variable value: C:\HJava



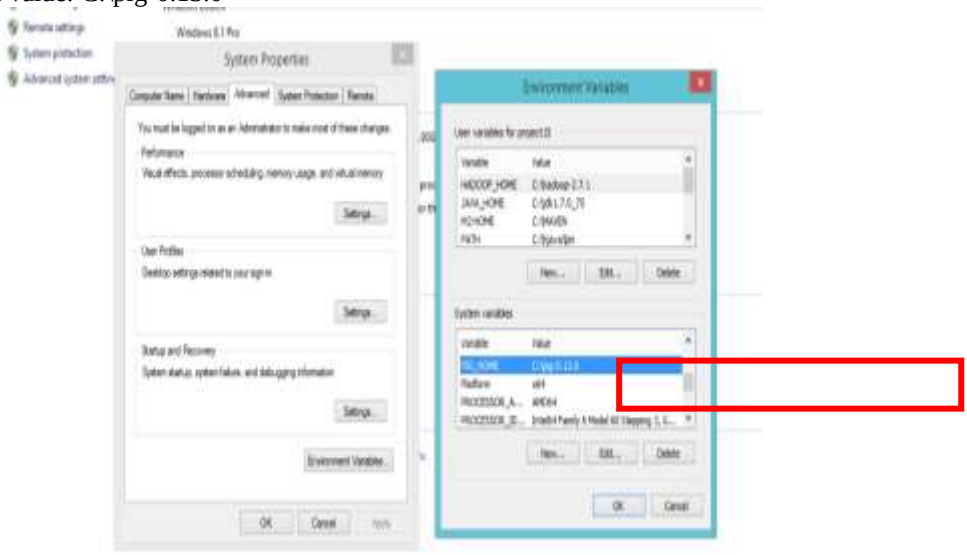
5. Now create new in system variables for *M2_HOME*:
   In variable name: *M2_HOME*
   Variable value: C:\maven



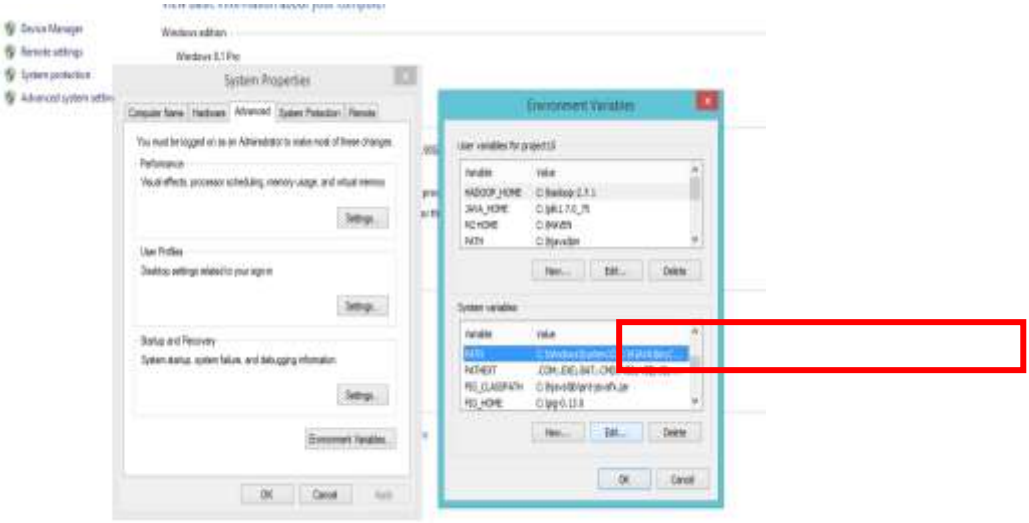
6. Now create new in system variables for *HIVE_HOME*:
   In variable name: *HIVE_HOME*
   Variable value: C:\apache-hive-1.2.1-src



7. Now create new in system variables for *PIG_HOME*:
   In variable name: *PIG_HOME*
   Variable value: C:\pig-0.13.0



8. Now Edit in system variables for *PATH variable*:
   In variable name: *PATH*
   Variable value: C:\Windows\System32;C:\HJAVA\bin;C:\MAVEN\bin;C:\protobuf-master;C:\hadoop-2.7.1\bin;C:\apache-hive-1.2.1-src\bin;C:\pig-0.13.0\bin; [4]



B. Deploying Hadoop:

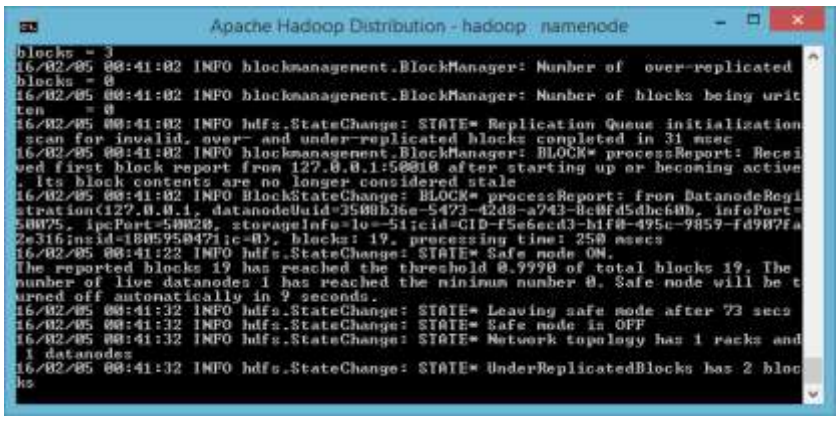
Once the configuration process is complete, it is time to start the Hadoop daemons. This is done by performing the following steps:

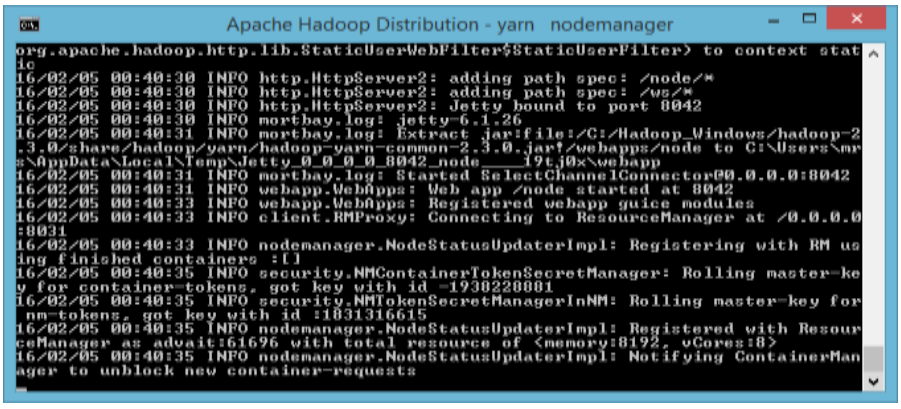
1. Before starting the daemons, we can format the NameNode by issuing the following command:

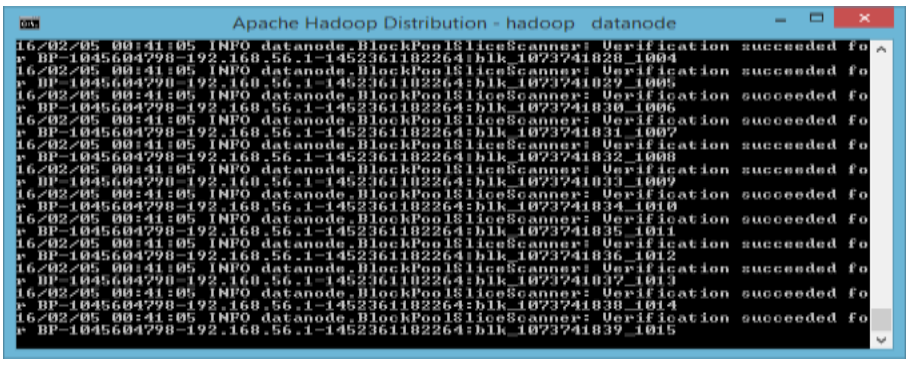
hdfs namenode –format

The Fig. 8 shows the screenshot which shows the output of the format command. Now the HDFS is formatted and ready to use. Since we have not specified a particular directory name, the NameNode creates the C:\hadoop directory to store all of the metadata. [1]

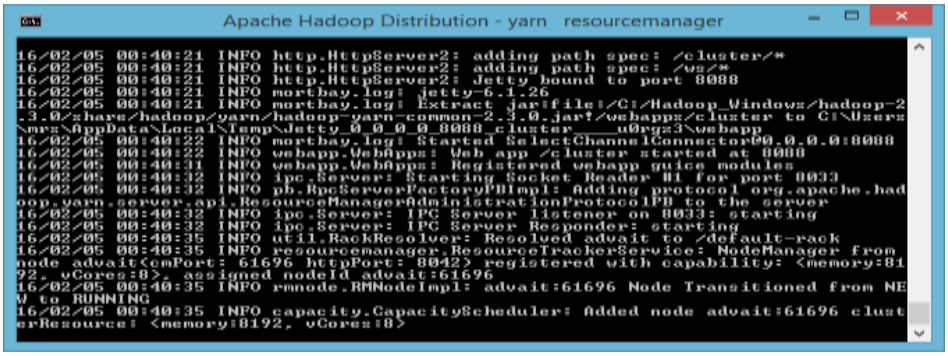
2. We then start the HDFS daemons, the NameNode, and the DataNode as shown in Fig. 9 and Fig. 10.This is done by issuing start-dfs.cmd. This command script is present in the %HADOOP_HOME%\sbin folder.







3. Next we need to start the YARN to run MapReduce jobs. This can be done by the start-yarn.cmd file present in the sbin folder. The ResourceManager and the NodeManager start in two separate command windows.



4. By navigating to http://localhost:50070 on the browser, the user should now be able to see the web endpoint for HDFS .It gives an overview of the health of HDFS and the different parameters that were used to configure it.

5. We can open the **Resource Manager** and **Node Manager** at http://localhost:8042

III. Languages: MAPREDUCE

MapReduce is intended to process expansive datasets for specific kinds of distributable issues. It endeavors to spread the work over an extensive number of hubs and enables those hubs to process the information in parallel. You can't include conditions inside the information, implying that you can't have a necessity that one record in a dataset must be prepared before another. Results from the underlying parallel handling are sent to extra hubs where the information is consolidated to take into consideration promote decreases of the information.[5]

The initial step is the guide step. It takes a subset of the full dataset called an info split and applies to each column in the information split a task that you have composed, for example, parsing each character string. The yield information is cradled in memory and spills to circle. It is arranged and apportioned by key utilizing the default partitioner. A consolidation sort sorts each segment. There might be numerous guide activities running in parallel with each other, every one handling an alternate information split. The segments are

rearranged among the reducers. For instance, parcel 1 goes to reducer 1. The second guide assignment likewise sends its parcel 1 to reducer 1. Segment 2 goes to another reducer.

IV. Conclusion

Hadoop has proved its ability by handling Big Data to deal with Unstructured Data very efficiently. This design which has shifted from scale up to scale out method has allowed the processing of distributed data to be accomplished in a manner that is cost effective. Hadoop is now readily available on Windows Platform and so there is no need of installing Linux virtual Machines on Windows. Since Majority of the users are comfortable with user friendly windows operating system, we decided to deploy hadoop for processing text data on this platform.

References

[1] https://www.safaribooksonline.com/library/view/hadoopessentials/9781784396688/ch02s05.html
[2] https://hadoop.apache.org/docs/r1.2.1/hdfs_de ign.html
[3] https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
[4] Tom White, Hadoop The Definitive guide, O'Reilly,Yahoo Press
[5] http://www.ibm.com/developerworks/cloud/library/clopenstack-deployhadoop/