**<http://www.srccodes.com/p/article/45/run-hadoop-wordcount-mapreduce-example-windows>**

**Hadoop Installation Error**

C:\Users\1989\Downloads\hadoop-2.7.1.tar.gz: Cannot create symbolic link C:\Users\1989\Downloads\hadoop\hadoop-2.7.1\lib\native\libhadoop.so

You may need to run WinRAR as administrator

! A required privilege is not held by the client.

C:\Users\1989\Downloads\hadoop-2.7.1.tar.gz: Cannot create symbolic link C:\Users\1989\Downloads\hadoop\hadoop-2.7.1\lib\native\libhdfs.so

You may need to run WinRAR as administrator

! A required privilege is not held by the client.

<http://stackoverflow.com/questions/12499851/why-hadoop-is-tightly-bound-to-linux>

https://wiki.apache.org/hadoop/Hadoop2OnWindows

Instead of going through the intricacies of deploying and configuring apache Hadoop on a Linux machine, to build a single node cluster, just for learning hadoop, I would suggest you rather go for a single node vm sandbox of cloudera or Horton works.  
They have centos vm I guess.  
[Cloudera VM Link](http://www.cloudera.com/content/cloudera/en/downloads/quickstart_vms/cdh-5-4-x.html" \t "https://www.quora.com/_blank)  
[HortonWorks VM Link](http://hortonworks.com/hdp/downloads/" \t "https://www.quora.com/_blank)

To run Hadoop on Windows you would need Hortonworks or Cloudera's CDH. Thought they may be easy to operate but you'll never find out how different frameworks integrate with each other within Hadoop as this would already have been done in Hortonworks or CDH.  
  
Whereas in Linux you can start from scratch, building everything and configuring everything on your own, this would enable you to grasp the concepts in a much more efficient manner.

#### Cygwin

As I said in the introduction, Hadoop assumes Linux (or a Unix flavor OS) is being used to run Hadoop. This assumption is buried pretty deeply. Various parts of H**adoop are executed using shell scripts that will only work on a Linux shell.** It also uses passwordless secure shell (SSH) to communicate between computers in the Hadoop cluster. T**he best way to do these things on Windows is to make Windows act more like Linux. You can do this using [Cygwin](http://www.cygwin.com/),** which provides a “Linux-like environment for Windows” that allows you to use Linux-style command line utilities as well as run really useful L**inux-centric software like OpenSSH.** Go [download the latest version of Cygwin](http://www.cygwin.com/). **Don’t install it yet.** I’ll describe how you need to install it below.

https://wiki.apache.org/hadoop/Hadoop2OnWindows

**I configure this setting**

<http://toodey.com/2015/08/10/hadoop-installation-on-windows-without-cygwin-in-10-mints/>

t has two main parts –

a data processing framework and  
a distributed filesystem for data storage.

a system like MapReduce to actually process the data.MapReduce runs as a series of jobs, with each job essentially a separate Java application that goes out into the data and starts pulling out information as needed. Using MapReduce instead of a query gives data seekers a lot of power and flexibility, but also adds a lot of complexity.

There are tools to make this easier: Hadoop includes [Hive](http://hive.apache.org/), another Apache application that helps convert query language into MapReduce jobs, for instance. But MapReduce’s complexity and its limitation to one-job-at-a-time batch processing tends to result in Hadoop getting used more often as a data warehousing than as a data analysis tool.

## Solution

**Storage:** This huge amount of data, Hadoop uses HDFS (Hadoop Distributed File System) which uses commodity hardware to form clusters and store data in a distributed fashion. It works on Write once, read many times principle.

**Processing:** Map Reduce paradigm is applied to data distributed over network to find the required output.

**Analyze:** Pig, Hive can be used to analyze the data.

**Cost:** Hadoop is open source so the cost is no more an issue.

**Environment required for Hadoop:** The production environment of Hadoop is UNIX, but it can also be used in Windows using Cygwin.

<https://prismoskills.appspot.com/lessons/System_Design_and_Big_Data/Chapter_01_-_Hadoop.jsp>

**Hive**  ()  provides SQL like querying capabilities to view data stored in the HDFS.  
It has its own query language called HiveQL.

**Pig**  ()  is a language used to run MapReduce jobs on Hadoop.  
It supports MapReduce programs in several languages including Java.

Diff PIG HIVE(yahoo vs fb) ==tool to make reducemap easier

<https://www.dezyre.com/article/difference-between-pig-and-hive-the-two-key-components-of-hadoop-ecosystem/79>

<https://www.youtube.com/watch?v=x-PCNX4prLA>