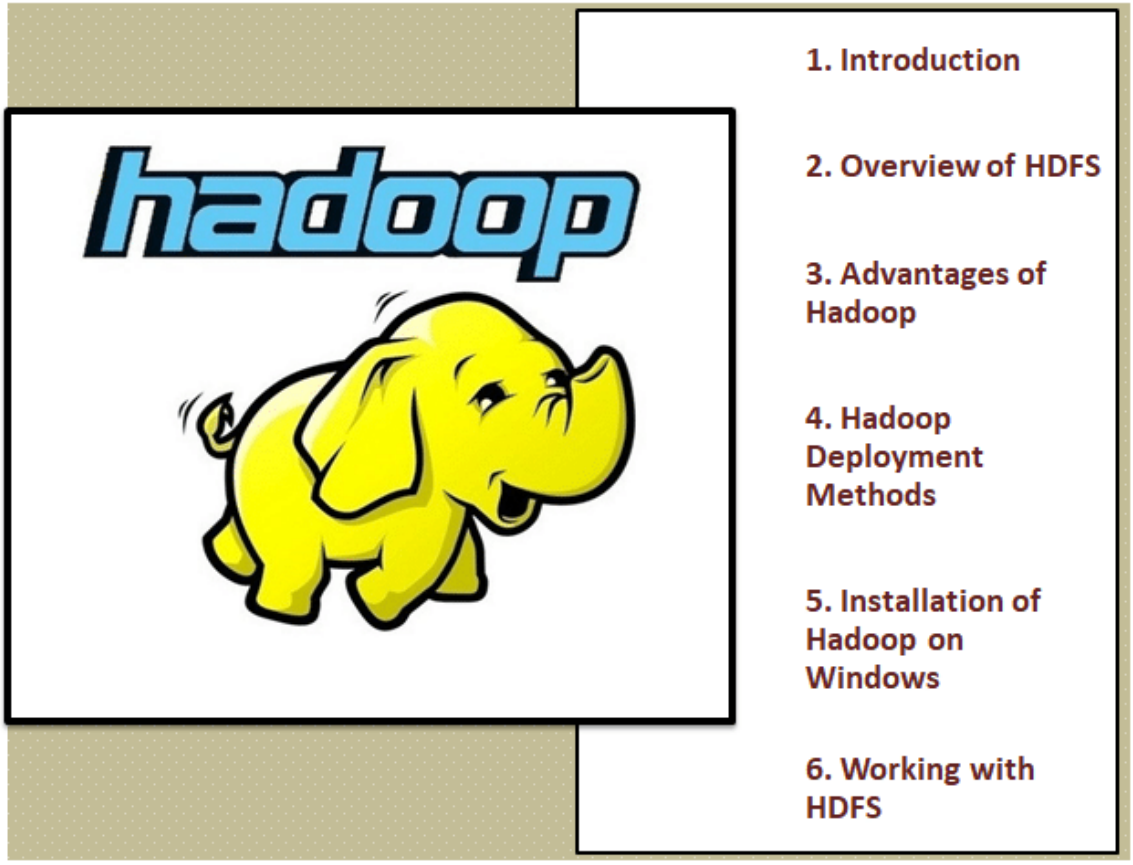
9923170071 / 8108094992 info@dimensionless.in

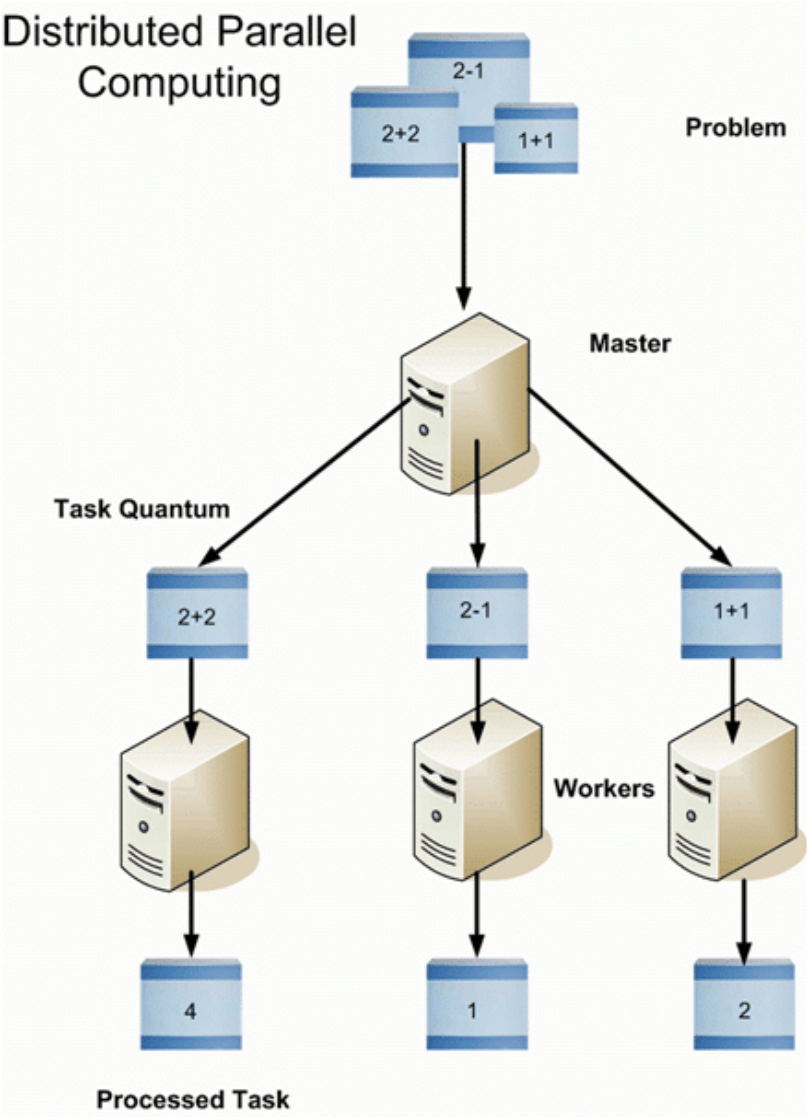
How to Install and Run Hadoop on Windows for Beginners

by Priyanka Gupta | Apr 9, 2019 | Big Data, Data Science, Hadoop | 7 comments



Introduction

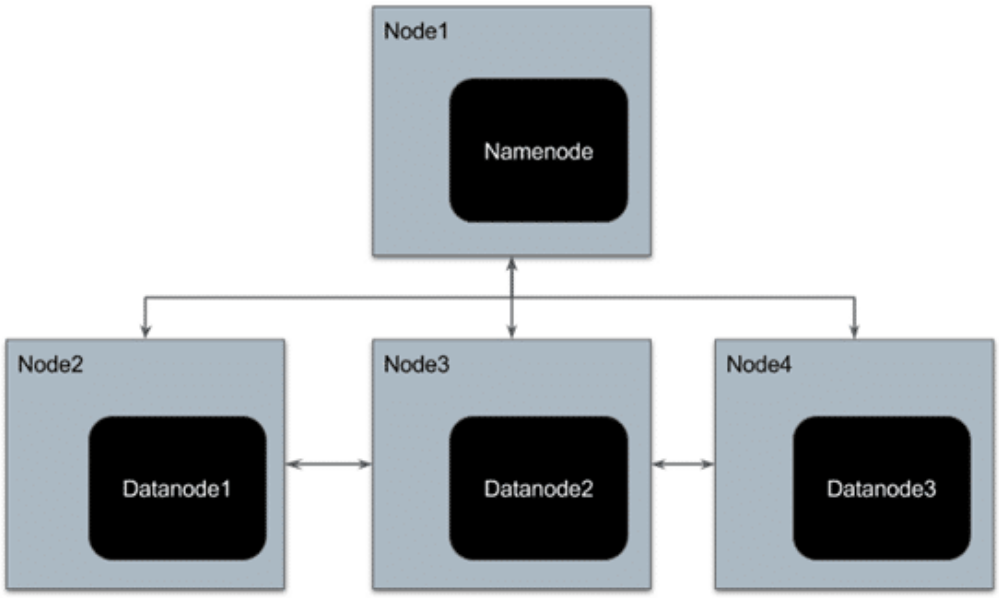
Hadoop is a software framework from Apache Software Foundation that is used to store and process Big Data. It has two main components; Hadoop Distributed File System (HDFS), its storage system and MapReduce, is its data processing framework. Hadoop has the capability to manage large datasets by distributing the dataset into smaller chunks across multiple machines and performing parallel computation on it .



Overview of HDFS

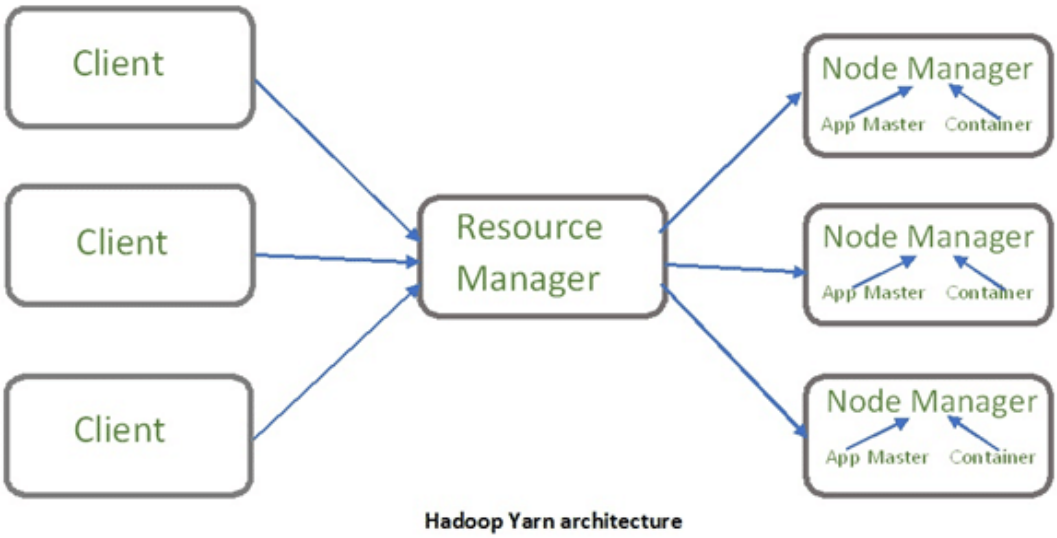
Hadoop is an essential component of the Big Data industry as it provides the most reliable storage layer, HDFS, which can scale massively. Companies like Yahoo and Facebook use HDFS to store their data.

HDFS has a master-slave architecture where the master node is called NameNode and slave node is called DataNode. The NameNode and its DataNodes form a cluster. NameNode acts like an instructor to DataNode while the DataNodes store the actual data.



source: Hasura

There is another component of Hadoop known as YARN. The idea of Yarn is to manage the resources and schedule/monitor jobs in Hadoop. Yarn has two main components, Resource Manager and Node Manager. The resource manager has the authority to allocate resources to various applications running in a cluster. The node manager is responsible for monitoring their resource usage (CPU, memory, disk) and reporting the same to the resource manager.



Hadoop Yarn architecture

source: GeeksforGeeks

To understand the Hadoop architecture in detail, refer this blog

Advantages of Hadoop

1. Economical – Hadoop is an open source Apache product, so it is free software. It has hardware cost associated with it. It is cost effective as it uses commodity hardware that are cheap machines to store its datasets and not any specialized machine.

2. Scalable – Hadoop distributes large data sets across multiple machines of a cluster. New machines can be easily added to the nodes of a cluster and can scale to thousands of nodes storing thousands of terabytes of data.

3. Fault Tolerance – Hadoop, by default, stores 3 replicas of data across the nodes of a cluster. So if any node goes down, data can be retrieved from other nodes.

4. Fast – Since Hadoop processes distributed data parallelly, it can process large data sets much faster than the traditional systems. It is highly suitable for batch processing of data.

5. Flexibility – Hadoop can store structured, semi-structured as well as unstructured data. It can accept data in the form of textfile, images, CSV files, XML files, emails, etc

6. Data Locality – Traditionally, to process the data, the data was fetched from the location it is stored, to the location where the application is submitted; however, in Hadoop, the processing application goes to the location of data to perform computation. This reduces the delay in processing of data.

7. Compatibility – Most of the emerging big data tools can be easily integrated with Hadoop like Spark. They use Hadoop as a storage platform and work as its processing system.

Hadoop Deployment Methods

1. Standalone Mode – It is the default mode of configuration of Hadoop. It doesn't use hdfs instead, it uses a local file system for both input and output. It is useful for debugging and testing.

2. Pseudo-Distributed Mode – It is also called a single node cluster where both NameNode and DataNode resides in the same machine. All the daemons run on the same machine in this mode. It produces a fully functioning cluster on a single machine.

3. Fully Distributed Mode – Hadoop runs on multiple nodes wherein there are separate nodes for master and slave daemons. The data is distributed among a cluster of machines providing a production environment.

Hadoop Installation on Windows 10

As a beginner, you might feel reluctant in performing cloud computing which requires subscriptions. While you can install a virtual machine as well in your system, it requires allocation of a large amount of RAM for it to function smoothly else it would hang constantly.

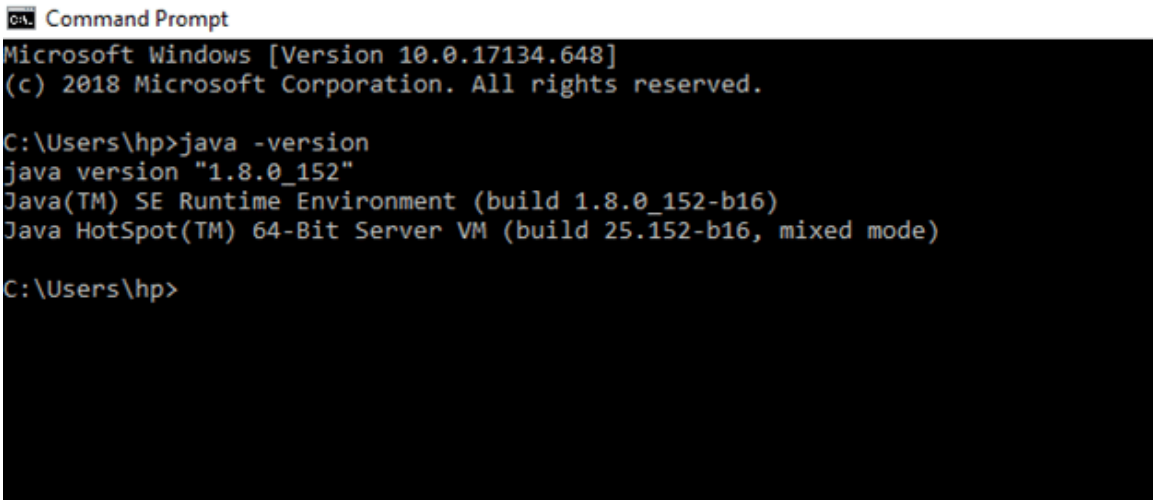
You can install Hadoop in your system as well which would be a feasible way to learn Hadoop.

We will be installing single node pseudo-distributed hadoop cluster on windows 10.

Prerequisite: To install Hadoop, you should have Java version 1.8 in your system.

Check your java version through this command on command prompt

```
java –version
```



If java is not installed in your system, then –

Go this link –

Accept the license,

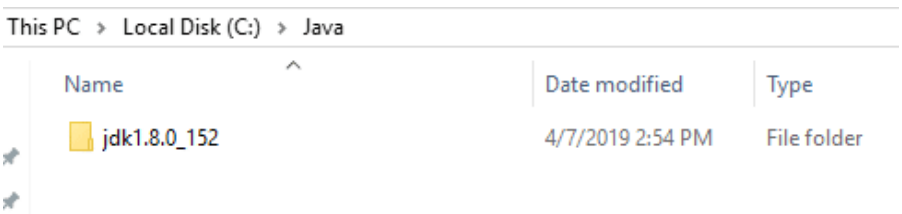
Java SE Development Kit 8u201 Demos and Samples Downloads

You must accept the Oracle BSD License. to download this software.

Accept License Agreement Decline License Agreement

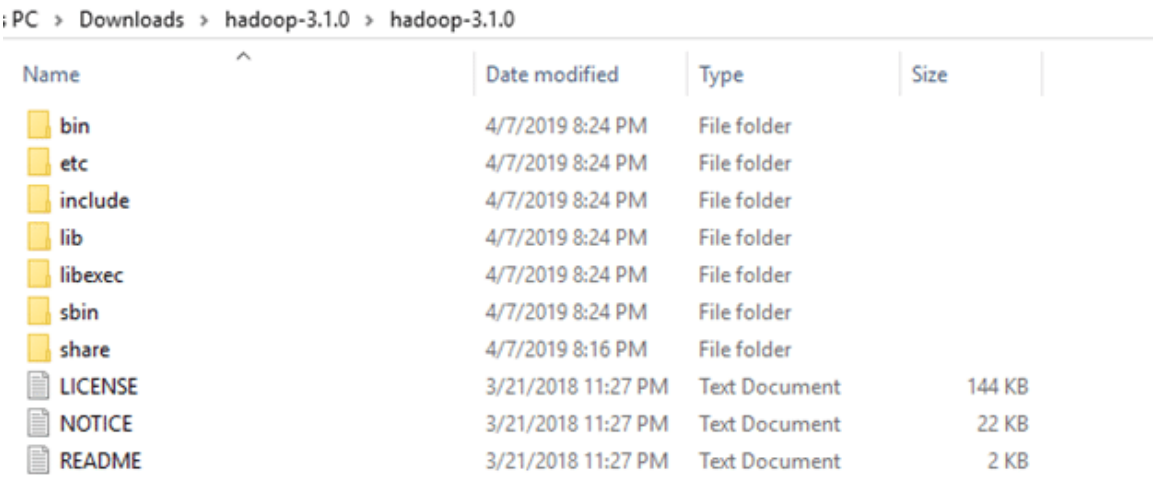
Product / File Description	File Size	Download
Linux ARM 32 Hard Float ABI	9.05 MB	jdk-8u201-linux-arm32-vfp-hflt-demos.tar.gz
Linux ARM 64 Hard Float ABI	9.06 MB	jdk-8u201-linux-arm64-vfp-hflt-demos.tar.gz
Linux x86	56.13 MB	jdk-8u201-linux-i586-demos.rpm
Linux x86	55.98 MB	jdk-8u201-linux-i586-demos.tar.gz
Linux x64	56.23 MB	jdk-8u201-linux-x64-demos.rpm
Linux x64	56.08 MB	jdk-8u201-linux-x64-demos.tar.gz
Mac OS X	56.25 MB	jdk-8u201-macosx-x86_64-demos.zip
Solaris SPARC 64-bit	12.2 MB	jdk-8u201-solaris-sparcv9-demos.tar.Z
Solaris SPARC 64-bit	8.46 MB	jdk-8u201-solaris-sparcv9-demos.tar.gz
Solaris x64	12.19 MB	jdk-8u201-solaris-x64-demos.tar.Z
Solaris x64	8.42 MB	jdk-8u201-solaris-x64-demos.tar.gz
Windows x86	56.96 MB	jdk-8u201-windows-i586-demos.zip
Windows x64	56.98 MB	jdk-8u201-windows-x64-demos.zip

Download the file according to your operating system. Keep the java folder directly under the local disk directory (C:\Java\jdk1.8.0_152) rather than in Program Files (C:\Program Files\Java\jdk1.8.0_152) as it can create errors afterwards.



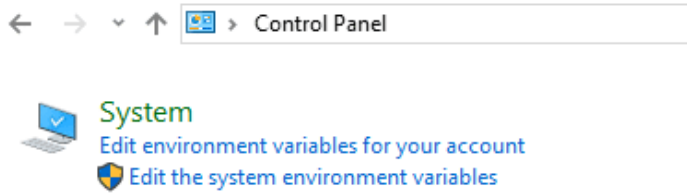
After downloading java version 1.8, download hadoop version 3.1 from this link –

Extract it to a folder.

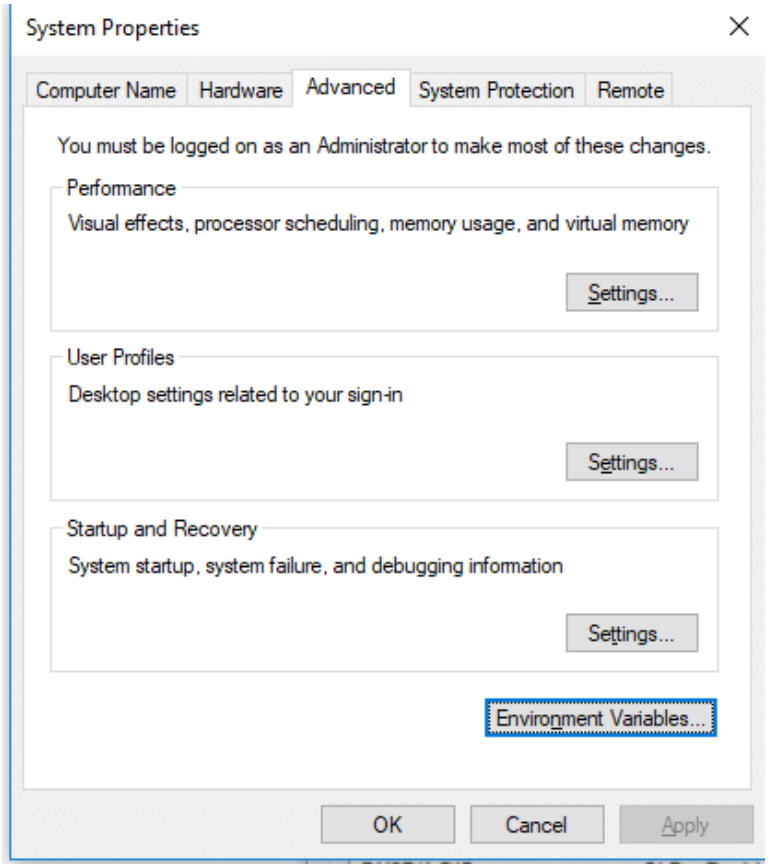


Setup System Environment Variables

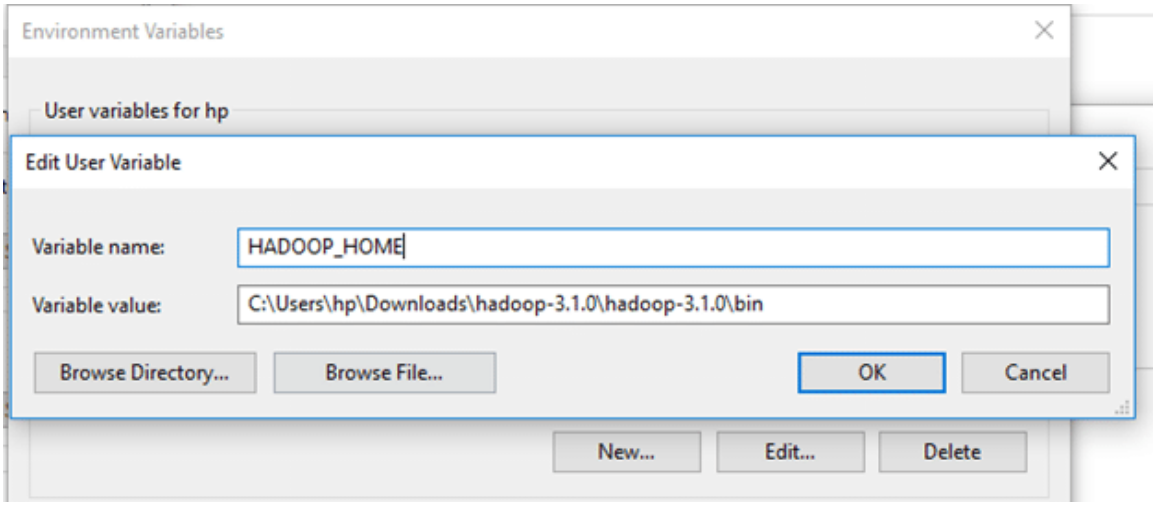
Open control panel to edit the system environment variable



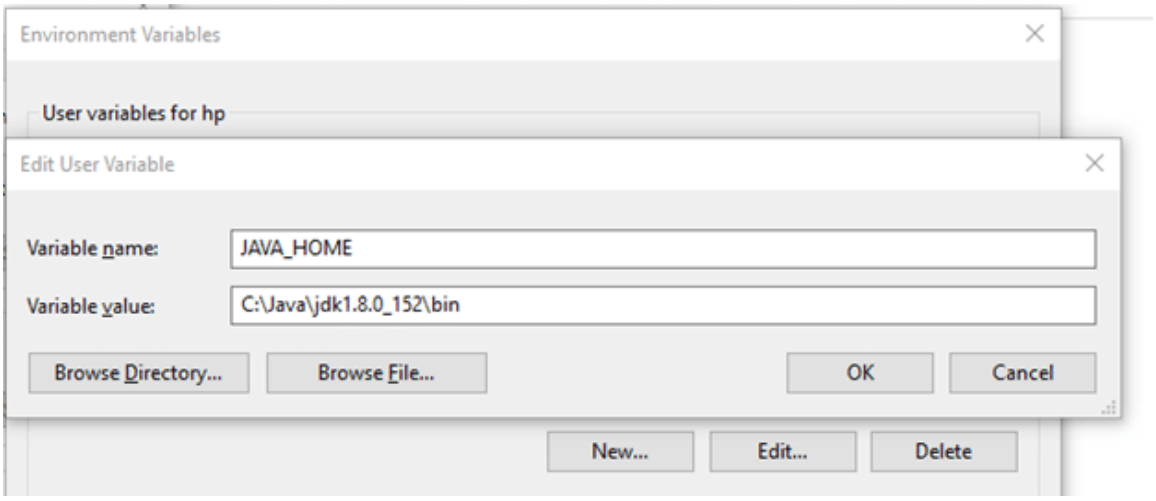
Go to environment variable in system properties



Create a new user variable. Put the Variable_name as HADOOP_HOME and Variable_value as the path of the bin folder where you extracted hadoop.

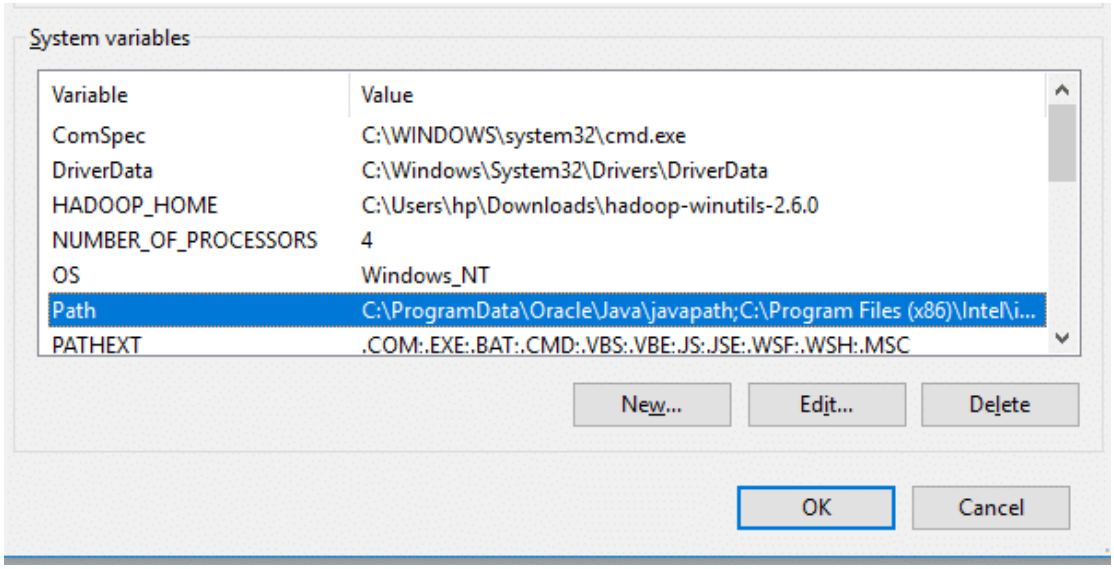


Likewise, create a new user variable with variable name as JAVA_HOME and variable value as the path of the bin folder in the Java directory.

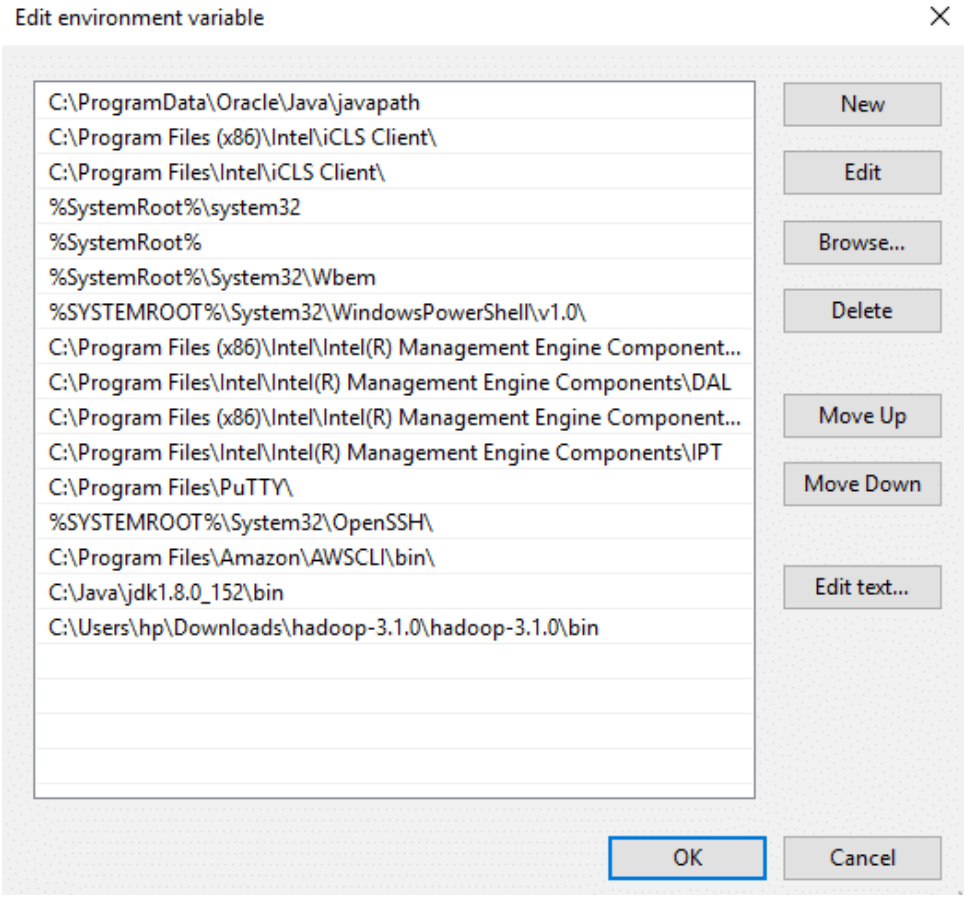


Now we need to set Hadoop bin directory and Java bin directory path in system variable path.

Edit Path in system variable

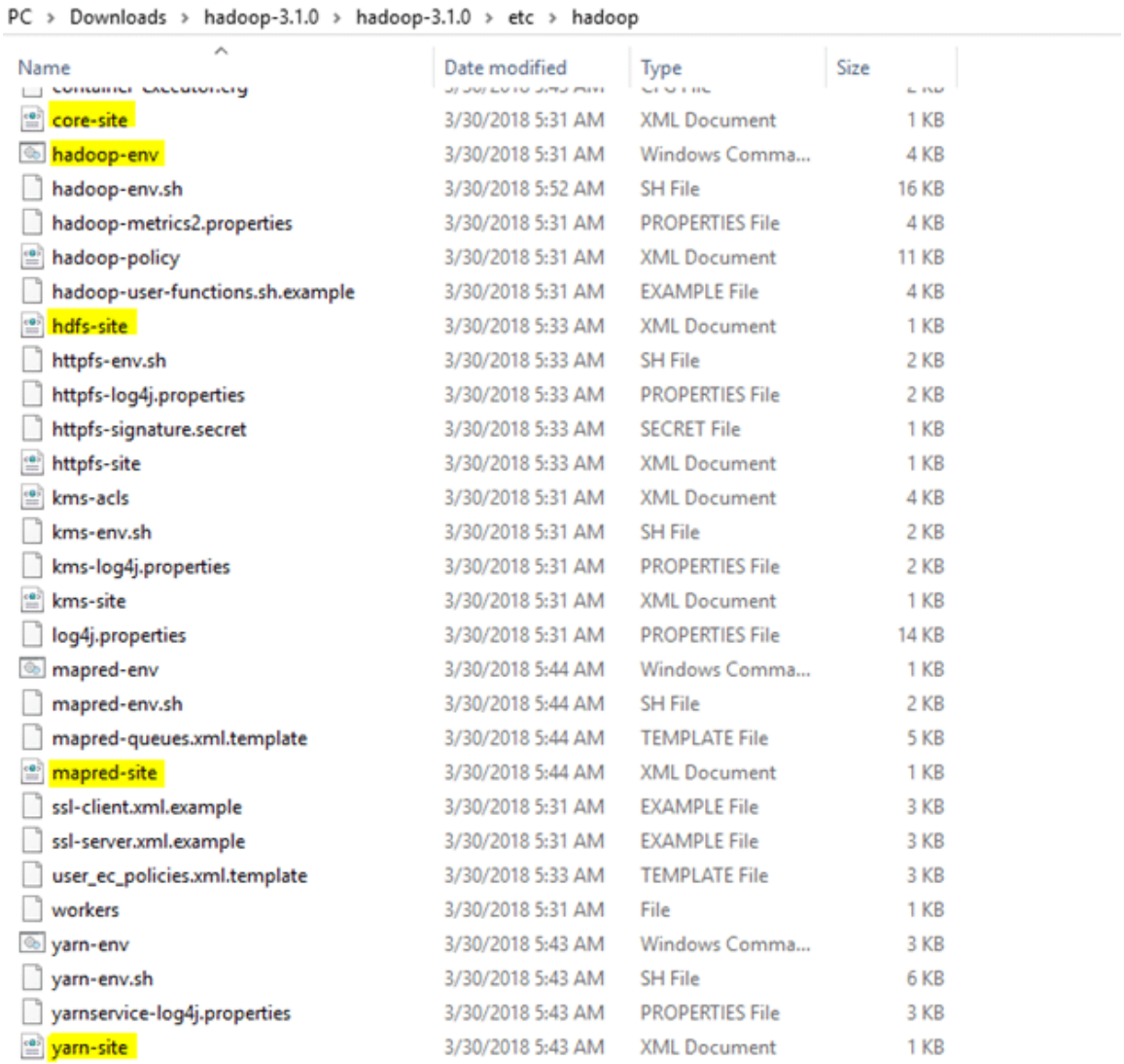


Click on New and add the bin directory path of Hadoop and Java in it.



Configurations

Now we need to edit some files located in the hadoop directory of the etc folder where we installed hadoop. The files that need to be edited have been highlighted.



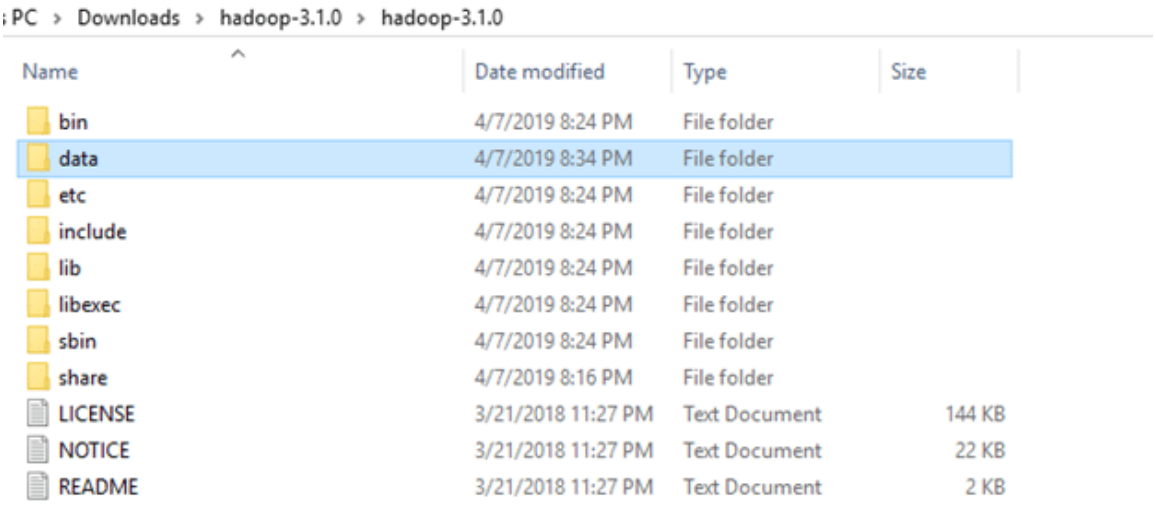
1. Edit the file core-site.xml in the hadoop directory. Copy this xml property in the configuration in the file

```
<configuration>
   <property>
       <name>fs.defaultFS</name>
       <value>hdfs://localhost:9000</value>
   </property>
</configuration>
```

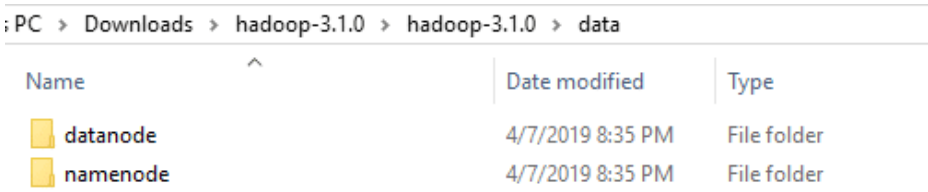
2. Edit mapred-site.xml and copy this property in the cofiguration

```
<configuration>
   <property>
       <name>mapreduce.framework.name</name>
       <value>yarn</value>
   </property>
</configuration>
```

3. Create a folder ‘data’ in the hadoop directory



Create a folder with the name ‘datanode’ and a folder ‘namenode’ in this data directory



4. Edit the file hdfs-site.xml and add below property in the configuration

Note: The path of namenode and datanode across value would be the path of the datanode and namenode folders you just created.

```
<configuration>
   <property>
        <name>dfs.replication</name>
        <value>1</value>
   </property>
   <property>
        <name>dfs.namenode.name.dir</name>
        <value>C:\Users\hp\Downloads\hadoop-3.1.0\hadoop-3.1.0\data\nameno
   </property>
   <property>
```

```
        <name>dfs.datanode.data.dir</name>
        <value> C:\Users\hp\Downloads\hadoop-3.1.0\hadoop-3.1.0\data\datan
    </property>
</configuration>
```

5. Edit the file yarn-site.xml and add below property in the configuration

```
<configuration>
    <property>
         <name>yarn.nodemanager.aux-services</name>
         <value>mapreduce_shuffle</value>
    </property>
    <property>
         <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
         <value>org.apache.hadoop.mapred.ShuffleHandler</value>
    </property>
</configuration>
```

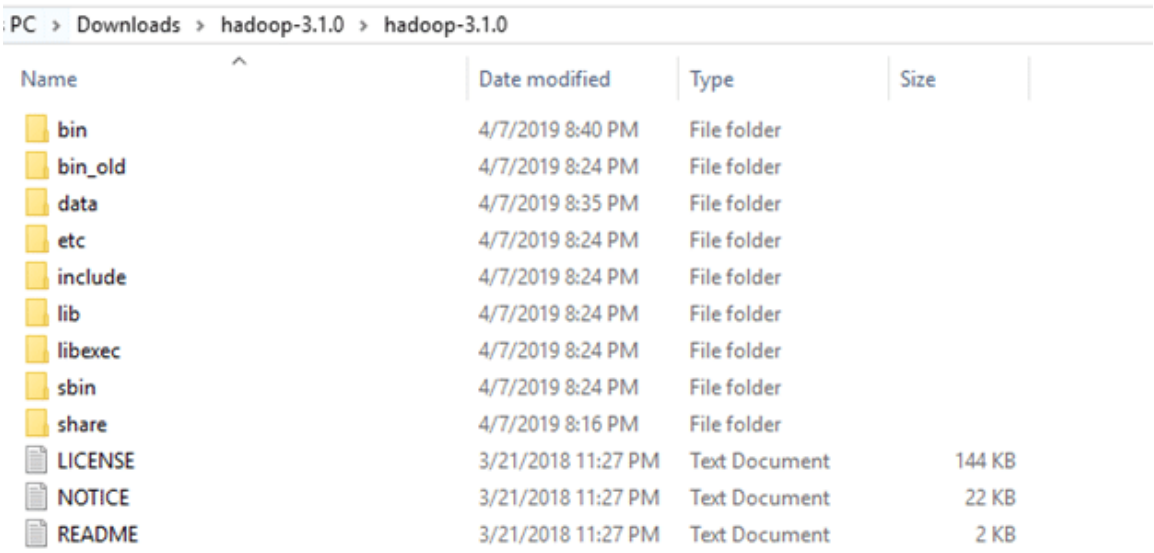
6. Edit hadoop-env.cmd and replace %JAVA_HOME% with the path of the java folder where your jdk 1.8 is installed

Hadoop needs windows OS specific files which does not come with default download of hadoop.

To include those files, replace the bin folder in hadoop directory with the bin folder provided in this github link.

https://github.com/s911415/apache-hadoop-3.1.0-winutils

Download it as zip file. Extract it and copy the bin folder in it. If you want to save the old bin folder, rename it like bin_old and paste the copied bin folder in that directory.



Check whether hadoop is successfully installed by running this command on cmd-

```
hadoop version
```

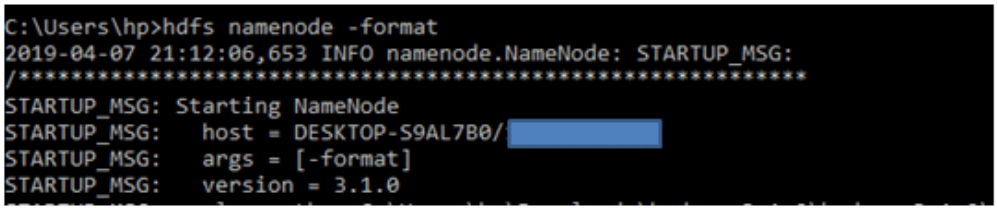
Since it doesn't throw error and successfully shows the hadoop version, that means hadoop is successfully installed in the system.

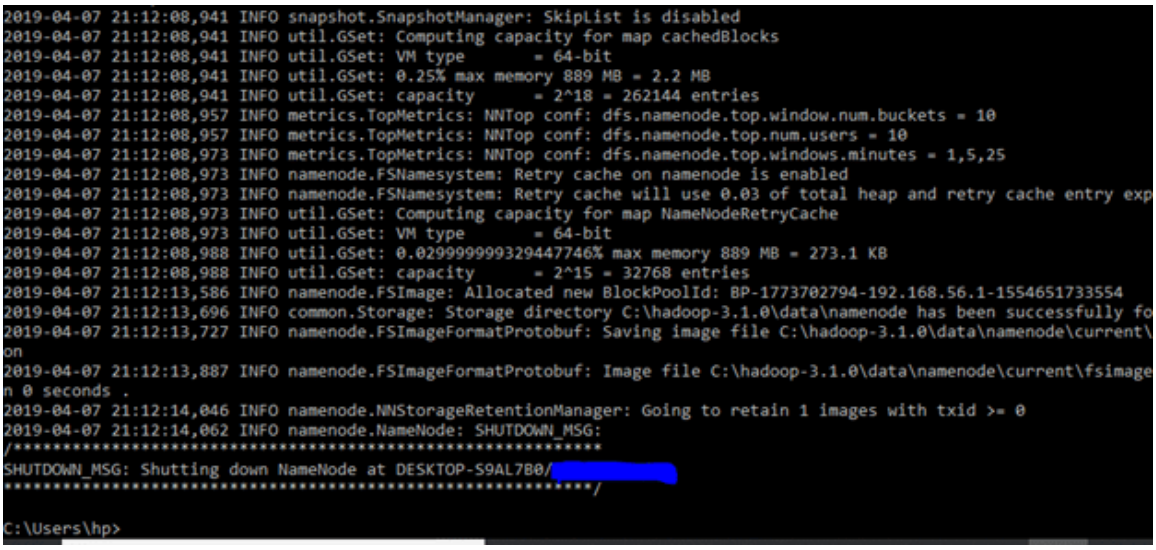
Format the NameNode

Formatting the NameNode is done once when hadoop is installed and not for running hadoop filesystem, else it will delete all the data inside HDFS. Run this command-

```
hdfs namenode –format
```

It would appear something like this –





Now change the directory in cmd to sbin folder of hadoop directory with this command,

(Note: Make sure you are writing the path as per your system)

```
cd C:\Users\hp\Downloads\hadoop-3.1.0\hadoop-3.1.0\sbin
```

Start namenode and datanode with this command –

```
start-dfs.cmd
```

```
C:\Users\hp>cd C:\Users\hp\Downloads\hadoop-3.1.0\hadoop-3.1.0\sbin

C:\Users\hp\Downloads\hadoop-3.1.0\hadoop-3.1.0\sbin>start-dfs.cmd

C:\Users\hp\Downloads\hadoop-3.1.0\hadoop-3.1.0\sbin>
```

Two more cmd windows will open for NameNode and DataNode

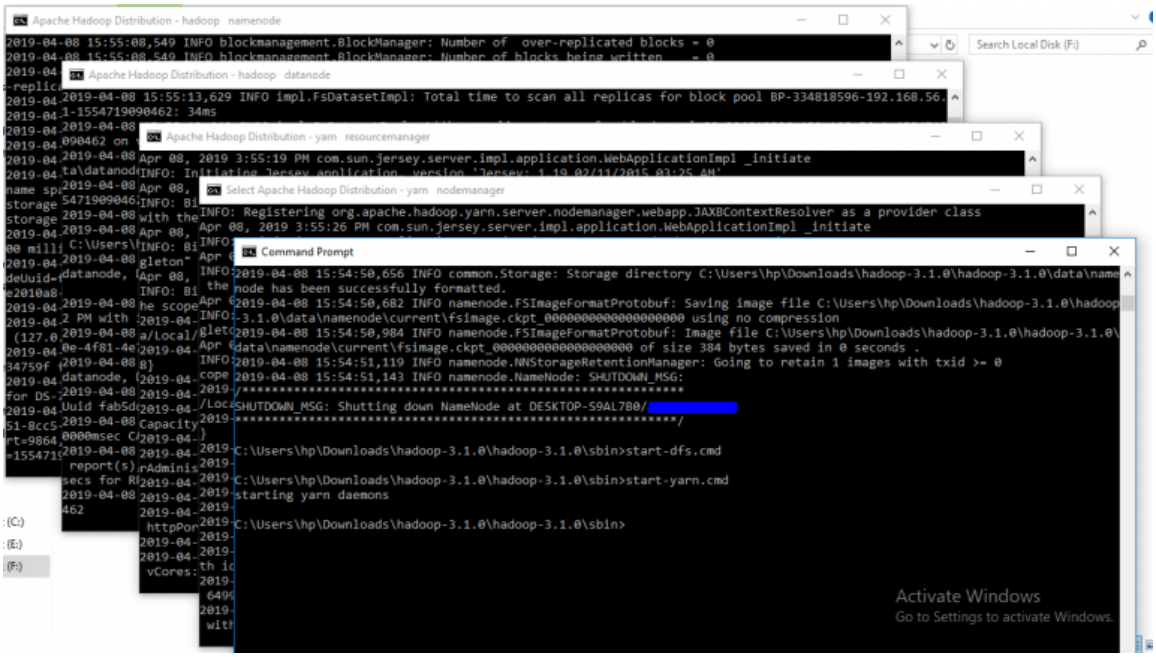
Now start yarn through this command-

`start-yarn.cmd`

```
C:\Users\hp\Downloads\hadoop-3.1.0\hadoop-3.1.0\sbin>start-yarn.cmd
starting yarn daemons

C:\Users\hp\Downloads\hadoop-3.1.0\hadoop-3.1.0\sbin>
```

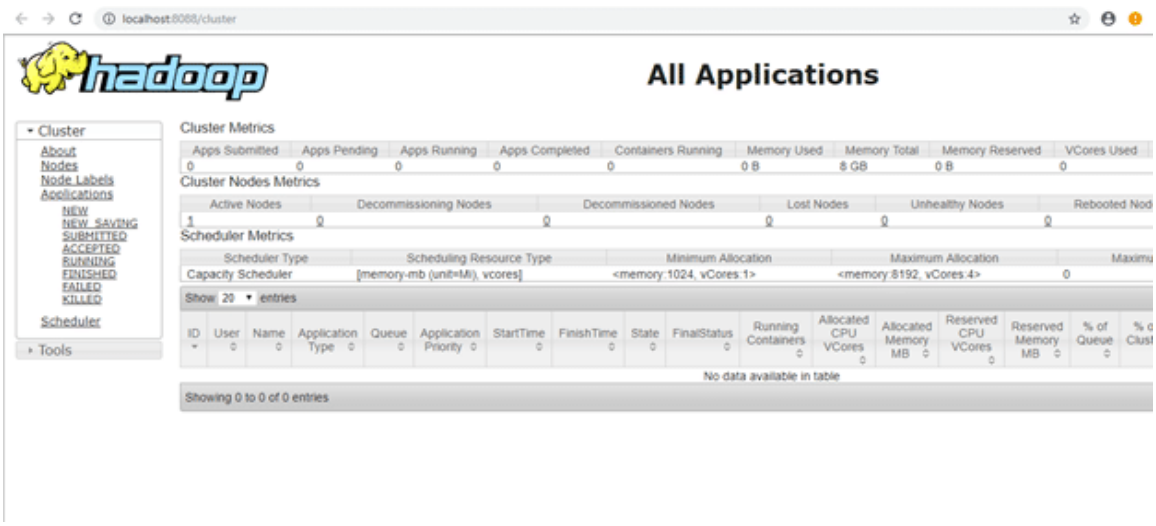
Two more windows will open, one for yarn resource manager and one for yarn node manager.



Note: Make sure all the 4 Apache Hadoop Distribution windows are up n running. If they are not running, you will see an error or a shutdown message. In that case, you need to debug the error.

To access information about resource manager current jobs, successful and failed jobs, go to this link in browser-

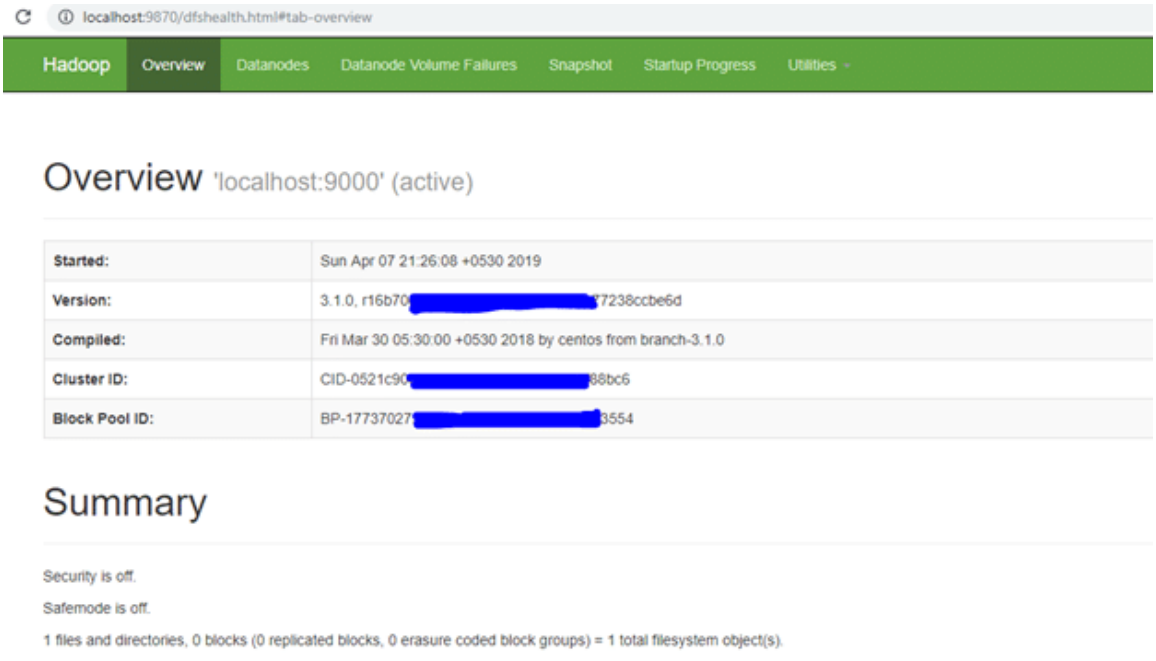
http://localhost:8088/cluster



To check the details about the hdfs (namenode and datanode),

Open this link on browser-

Note: If you are using Hadoop version prior to 3.0.0 – Alpha 1, then use port here



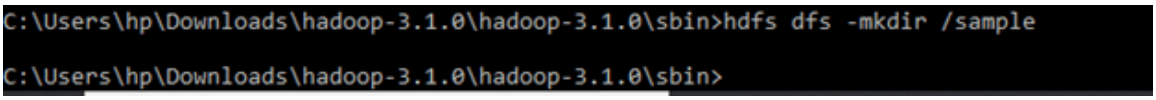
Working with HDFS

I will be using a small text file in my local file system. To put it in hdfs using hdfs command line tool.

I will create a directory named ‘sample’ in my hadoop directory using the

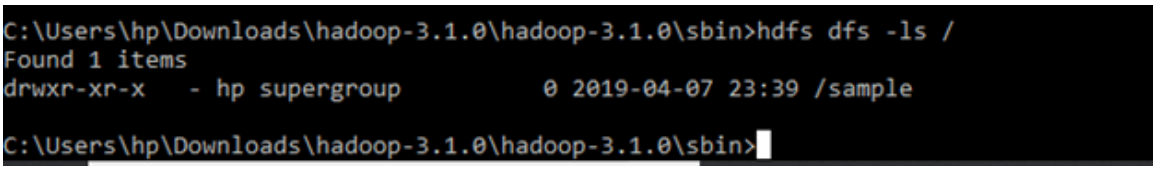
following command-

```
hdfs dfs –mkdir /sample
```



To verify if the directory is created in hdfs, we will use 'ls' command which will list the files present in hdfs –

```
hdfs dfs –ls /
```



Then I will copy a text file named 'potatoes' from my local file system to this folder that I just created in hdfs using copyFromLocal command-

```
hdfs dfs –copyFromLocal C:\Users\hp\Downloads\potatoes.txt /sample
```

To verify if the file is copied to the folder, I will use 'ls' command by specifying the folder name which will read the list of files in that folder –

```
hdfs dfs –ls /sample
```

To view the contents of the file we copied, I will use cat command-

```
hdfs dfs -cat /sample/potatoes.txt
```

To Copy file from hdfs to local directory, I will use get command –

```
hdfs dfs -get /sample/potatoes.txt C:\Users\hp\Desktop\priyanka
```

These were some basic hadoop commands. You can refer to this HDFS commands guide to learn more here

Conclusion

Hadoop MapReduce can be used to perform data processing activity. However, it possessed limitations due to which frameworks like Spark and Pig emerged and have gained popularity. A 200 lines of MapReduce code can be written with less than 10 lines of Pig code. Hadoop has various other components in its ecosystem like Hive, Sqoop, Oozie, and HBase. You can download this software as well in your windows system to perform data processing operations using cmd.

Follow this link, if you are looking to **learn more about data science online!**

You can **follow this link for our Big Data course!**

Additionally, if you are having an interest in **learning Data Science, click here to start Online Data Science Courses**

7 Comments

routeripaddress.site on June 2, 2019 at 11:09 am

Appreciating the time and energy you put into your site and in depth information you
provide. It's nice to come across a blog every once in a while that isn't the same outdated rehashed material.
Great read! I've saved your site and I'm adding your RSS feeds to my Google
account.

Reply

pilgrim on July 23, 2019 at 10:17 am

This site certainly has all the information and facts I needed about this subject and didn't know who to ask.

Reply

spyglass on July 27, 2019 at 8:33 pm

What's up mates, good piece of writing and nice arguments commented here, I am really enjoying by these.

Reply

seo tools vps on July 30, 2019 at 11:20 pm

Thanks for ones marvelous posting! I quite enjoyed reading it, you can be a great author.I will make sure to bookmark your blog and
may come back at some point. I want to encourage you to definitely continue your great writing,
have a nice weekend!

Reply

situs bandarqq on August 10, 2019 at 8:29 am

No matter if some one searches for his vital thing,
so he/she wants to be available that in detail, thus that thing is maintained over here.

Reply

온라인카지노 on August 21, 2019 at 6:26 am

Definitely consider that that you stated. Your favourite justification seemed
to be on the web the easiest factor to consider of.
I say to you, I definitely get annoyed at the same time as folks consider issues that they just do not know about.
You managed to hit the nail upon the top and defined out the whole thing without having side-effects , other people can take a signal.
Will likely be again to get more. Thanks

Reply

Joette Daras on August 22, 2019 at 3:16 am

I am a mother and this helped me!

Reply

★★★★★ A very big thank you to you all sharing her good work as an expert in crypto and forex trade option. Thanks for... read more

Deepak Prasad
2 years ago
★☆☆☆☆ Faculty knowledge is good but they didn't cover most of the topics which was mentioned in curriculum during online... read more

Ritika Khandelwal
3 years ago
★★★★★ Dimensionless is great place for you to begin exploring Data science under the guidance of experts. Both Himanshu and... read more

Rupal Gupta
4 years ago
★★★★★ My experience with the data science course at Dimensionless has been extremely positive. The course was effectively... read more

Durgesh Tiwari
4 years ago
★★★★★ It has been a great experience with Dimensionless . Especially from the support team , once you get enrolled , you... read more

Jasminder Singh
4 years ago
★★★★★ The training experience has been really good! Specially the support after training!! HR team is really good. They keep... read more

Next Reviews

Recent Posts

My Journey: From Business Analyst to Data Scientist

Test Engineer to Data Science: Career Switch

Data Engineer to Data Scientist : Career Switch

Data Science Trend Forecast: 2020

Best of 2019 Trends in Data Science

Topics

Agent-Based Modelling (1)

AI (1)

Analytics (18)

Artificial Intelligence (2)

Uncategorized (12)

Visualisation (9)

Tags

2019 2019 trends AI Analytics AWS Big Data BlockChain Business Analysts Career Switch Data Science Career Transition Career Transitions Cloud Computing clustering Data Mining Data Science data science careers data science courses data science job data science professional data science skillset Data Science Success Story data scientist data visualisation tools Deep Learning Dimensionless Technologies future ready careers Gradient boosting Hardware Engineer High Demand Kafka KNN Learn Data Science and Business Analytics Machine Learning Network Engineer nlp programming Python R random forest Research Associate Software Spark SVM trends visualization

FOR BUSINESSES

Corporate Training

SUBCRIBE FORM

Registration for Demo

ABOUT US

Baggage AI

COURSES

PG Diploma in Data Science and Artificial Intelligence

TCS iON ProCert – Artificial Intelligence Certification

Artificial Intelligence (AI) Specialization Program

Data Science Specialization Program

Data Analytics Program

Machine Learning Specialization

Tableau – Desktop Certified Associate Training | Dimensionless

Big Data Analytics

Deep Learning

KNOWLEDGE PORTAL

Blog

Videos

CONTACT US

+31 6 82109716

3047 Spring Hill Pkwy SE, Smyrna, GA 30080, United States

info@dimensionless.in

CONTACT US

+1 (470) 301 2795

info@dimensionless.in

CONTACT US

+31 6 82109716

Generaal Stedmanstraat 2505623 HZ Eindhoven, Netherlands

info@dimensionless.in

CONTACT US

+91 91378 44672

CIBA, 6th Floor, Agnel Technical Complex,Sector 9A,, Vashi, Navi Mumbai, Mumbai, Maharashtra 400703

B303, Sai Silicon Valley, Balewadi, Pune, Maharashtra 411045

info@dimensionless.in

POLICIES AND COPYRIGHT