

TABLE OF CONTENTS

Introduction	1
How to Install	2
Images/Screenshots/Installation	3
Importance of Hadoop	4
Stored Data to Data Node	5

INTRODUCTION

Apache Hadoop is a collection of open-source software utilities that facilitates using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model. Hadoop was originally designed for computer clusters built from commodity hardware, which is still the common use. It has since also found use on clusters of higher-end hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework.

The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model. Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packaged code into nodes to process the data in parallel. This approach takes advantage of data locality, where nodes manipulate the data they have access to. This allows the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking.

HOW TO INSTALL HADOOP

Step 1: go to <https://hadoop.apache.org/releases.html> and click 3.3.0 Binary Download

Step 2: Click any of the available mirror links to download Hadoop

Step 3: Make sure you have Java Developer Kit (JDK) installed. If it is not installed, go to <https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html>.

Step 4: Create a new Environment Variable. You can input any name on Variable Name, input the path of your folder in Variable Value, and copy the file path into System Variable.

Step 5: Open CMD and enter the following command: `cd Program Files\Java\jdk-13.0.2\bin`. Afterwards, type `javac`.

Step 6: After installing JDK, you may proceed to install Hadoop. Simply extract the files outside the archive and transfer them to Local C.

Step 7: After extracting the files, find the following files in the etc folder:

`-- core-site-- xml mode`

`-- hadoop-env-- cmd mode`

`-- mapred-site -- xml mode`

`-- yarn site -- xml mode`

`-- hdfs site - xml mode`

Step 8: Create a folder named “data” in Hadoop. In data, create two folders, datanode and namenode.

IMAGES/SCREENSHOTS/INSTALLATION

```
Select C:\Windows\system32\cmd.exe
```


-archives xarchhive1,...> specify a comma-separated list of archives to be unarchived on the compute machines

The general command line syntax is:
command [genericOptions] [commandOptions]

```
C:\(Users\R3)sh> cd Apache Hadoop Distribution -hadoop namenode  
e.hadoop.hdfs.web.resources.XAttrSetFlagParam.org.apache.hadoop.hdfs.web.resources.SnapshotNameParam.org.apache.hadoop.hf  
C:\(Users\R3)sh> e.hadoop.hdfs.web.resources.OldSnapshotNameParam.org.apache.hadoop.hdfs.web.resources.ExcludeDataNodesParam.org.apac  
web.resources.CreateFlagParam.org.apache.hadoop.hdfs.web.resources.NoRedirectParam.org.apache.hadoop.hdfs.web.resourc  
C:\(Users\R3)sh> ha.StoragePolicyParam.org.apache.hadoop.hdfs.web.resources.ECPolicyParam.org.apache.hadoop.hdfs.web.resources.NameSpaceQuo  
DEPRECATED: UstParam.org.apache.hadoop.hdfs.web.resources.StorageSpaceQuotaParam.org.apache.hadoop.hdfs.web.resources.StorageTypePar  
Instead use thm) throws java.io.IOException,java.lang.InterruptedExceptio, with URI template, "/" , is treated as a resource method  
#include <std WARNING: A sub-resource method, public javax.us.rs.core.Response org.apache.hadoop.hdfs.server.namenode.web.resources.  
#include <conNamenodeWebHdfsMethods.postRoot(org.apache.hadoop.security.UserGroupInformation.org.apache.hadoop.hdfs.web.resources.Del  
#include <striegationParam.org.apache.hadoop.hdfs.web.resources.UserParam.org.apache.hadoop.hdfs.web.resources.ConcatSourcesParam.org.apach  
#include <timestep.hadoop.hdfs.web.resources.PostOpParam.org.apache.hadoop.hdfs.web.resources.ExcludeDataNodesParam.org.apache.hadoop.hdfs.web.r  
#include <grapesources.BufferSizeParam.org.apache.hadoop.hdfs.web.resources.ExcludeDataNodesParam.org.apache.hadoop.hdfs.web.resourc  
.NewLengthParam.org.apache.hadoop.hdfs.web.resources.NoRedirectParam) throws java.io.IOException,java.lang.InterruptedEx  
void reg(); ception, with URI template, "/" , is treated as a resource method  
void register WARNING: A sub-resource method, public javax.us.rs.core.Response org.apache.hadoop.hdfs.server.namenode.web.resources.t  
void features(NamenodeWebHdfsMethods.deleteRoot(org.apache.hadoop.security.UserGroupInformation.org.apache.hadoop.hdfs.web.resourc  
void display(allegationParam.org.apache.hadoop.hdfs.web.resources.UserParam.org.apache.hadoop.hdfs.web.resources.DoAsParam.org.apache  
void show() : hadoop.hdfs.web.resources.DeleteOpParam.org.apache.hadoop.hdfs.web.resources.RecursiveParam.org.apache.hadoop.hdfs.web.r  
void search(roesources.SnapsShotNameParam) throws java.io.IOException,java.lang.InterruptedExceptio, with URI template, "/" , is treat  
void file deleted as a resource method  
void box(); 2020-12-21 16:13:51,330 INFO namenode.FSEditLog: Number of transactions: 2 Total time for transactions(ms): 47 Number of  
void features( transactions batched in Syncrs: 0 Number of syncrs: 2 SyncTimes(ms): 155  
void menu()); 2020-12-21 16:15:37,445 INFO namenode.FSEditLog: Number of transactions: 3 Total time for transactions(ms): 49 Number of  
transactions batched in Syncrs: 0 Number of syncrs: 3 SyncTimes(ms): 203  
struct per 2020-12-21 16:15:37,537 INFO hdfs.StatsChange: BLOCK* allocate blk_1073741825_1001, replicas=127.0.0.1:19866 for /robin_d  
{  
ir/HTML.C._COPYING  
char n2020-12-21 16:15:37,873 INFO namenode.FSNamesystem: BLOCK* blk_1073741825_1001 is COMMITTED but not COMPLETE(numNodes = 0  
r r< minimum + 1) In file /robin_dir/HTML.C._COPYING  
char p2020-12-21 16:15:39,284 INFO hdfs.StatsChange: DIR* completefile: /robin_dir/HTML.C._COPYING_is closed by DFSClient_NOM  
char gMAPREDUCE_448539138_1  
// time_t  
// time(&current);  
};  
  
(int main()  
  
menu();  
getch());
```

Type here to search

4:22 pm
12/21/2020



All Applications

Cluster

[About](#)
[Nodes](#)
[Node Labels](#)
[Applications](#)

NEW
NEW SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED

[Scheduler](#)

Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used
0	0	0	0	0	0 B

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost
1	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus
No data available in table											

Showing 0 to 0 of 0 entries

IMPORTANCE OF HADOOP

Why is Hadoop important?

- **Ability to store and process huge amounts of any kind of data, quickly.** With data volumes and varieties constantly increasing, especially from social media and the [Internet of Things \(IoT\)](#), that's a key consideration.
- **Computing power.** Hadoop's distributed computing model processes [big data](#) fast. The more computing nodes you use, the more processing power you have.
- **Fault tolerance.** Data and application processing are protected against hardware failure. If a node goes down, jobs are automatically redirected to other nodes to make sure the distributed computing does not fail. Multiple copies of all data are stored automatically.
- **Flexibility.** Unlike traditional relational databases, you don't have to preprocess data before storing it. You can store as much data as you want and decide how to use it later. That includes unstructured data like text, images and videos.
- **Low cost.** The open-source framework is free and uses commodity hardware to store large quantities of data.
- **Scalability.** You can easily grow your system to handle more data simply by adding nodes. Little administration is required.

STORING DATA IN DATA NODE

Step 1: Open CMD

Step 2: Create a file directory inside datanode by enter the command:

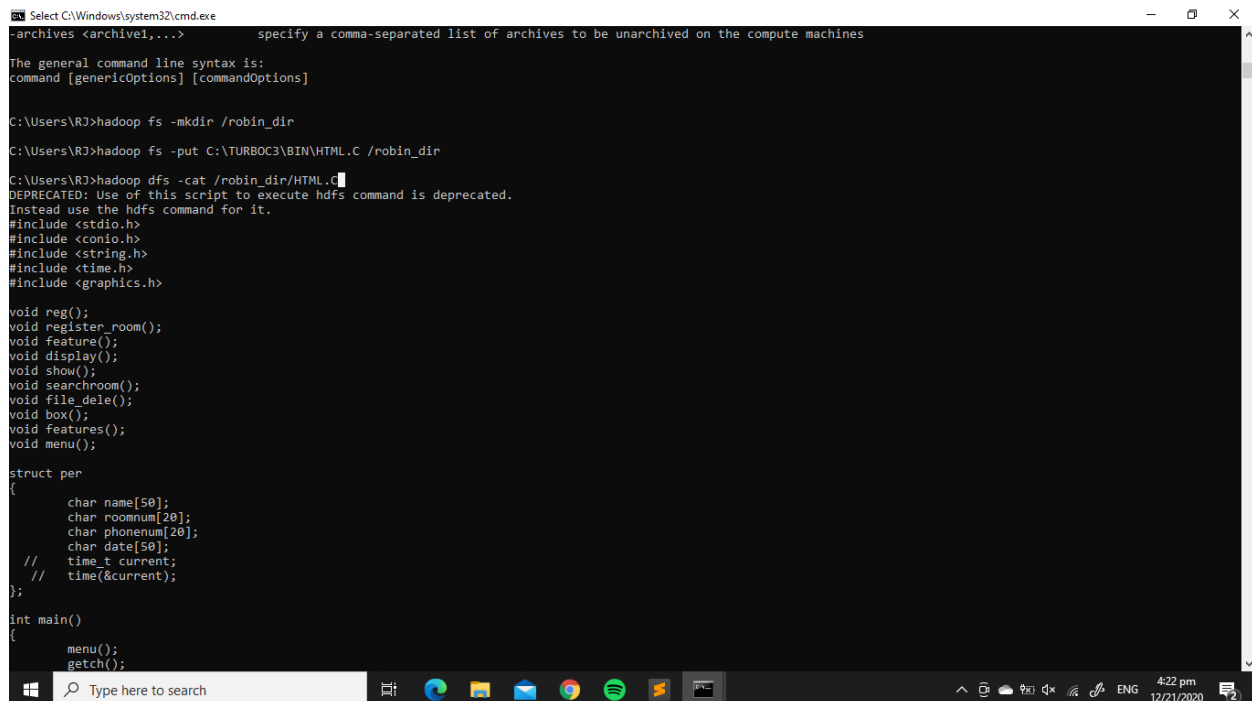
```
C:\Users\RJ>hadoop fs -mkdir /robin_dir
```

Step 3: Insert a file into robin_dir (or whatever you named it) using the command:

```
C:\Users\RJ>hadoop fs -put [Directory of the file you want to insert] /robin_dir
```

Step 4: After the file has been successfully inserted, you can have it displayed using this command:

```
C:\Users\RJ>hadoop fs -cat /robin_dir/[Name of the inserted file]
```



```

Select C:\Windows\system32\cmd.exe
-archives <archive1,...>      specify a comma-separated list of archives to be unarchived on the compute machines

The general command line syntax is:
command [genericOptions] [commandOptions]

C:\Users\RJ>hadoop fs -mkdir /robin_dir
C:\Users\RJ>hadoop fs -put C:\TURBOC3\BIN\HTML.C /robin_dir
C:\Users\RJ>hadoop dfs -cat /robin_dir/HTML.C
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
#include <stdio.h>
#include <conio.h>
#include <string.h>
#include <time.h>
#include <graphics.h>

void reg();
void register_room();
void feature();
void display();
void show();
void searchroom();
void file_dele();
void box();
void features();
void menu();

struct per
{
    char name[50];
    char roomnum[20];
    char phonenumber[20];
    char date[50];
    // time_t current;
    // time(&current);
};

int main()
{
    menu();
    getch();
}
```

Browse Directory

/

Show 25 entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	drwxr-xr-x	RJ	supergroup	0 B	Dec 21 16:15	0	0 B	robin_dir	

Showing 1 to 1 of 1 entries

Previous 1 Next

Hadoop, 2020.