

Name: Harsh Chheda
Roll Number: 22-15405

MSC COMPUTER SCIENCE

Subject: Big Data

Name: Harsh Chheda
Roll Number: 22-15405
Class: Msc. Computer Science (Part 2)
Subject: Big Data
Year: 2022-23

Name: Harsh Chheda
Roll Number: 22-15405

MSC COMPUTER SCIENCE

Subject: Big Data

INDEX			
NO	TITLE	PAGE NO	SIGN
1	Mongo DB Basic Commands	03 - 18	
2	Installation Of Hadoop	19 - 36	
3	Write a Hadoop MapReduce Program in Python.	37 - 45	

Practical 1

Q1) Show existing Databases and create a new DB MSC.



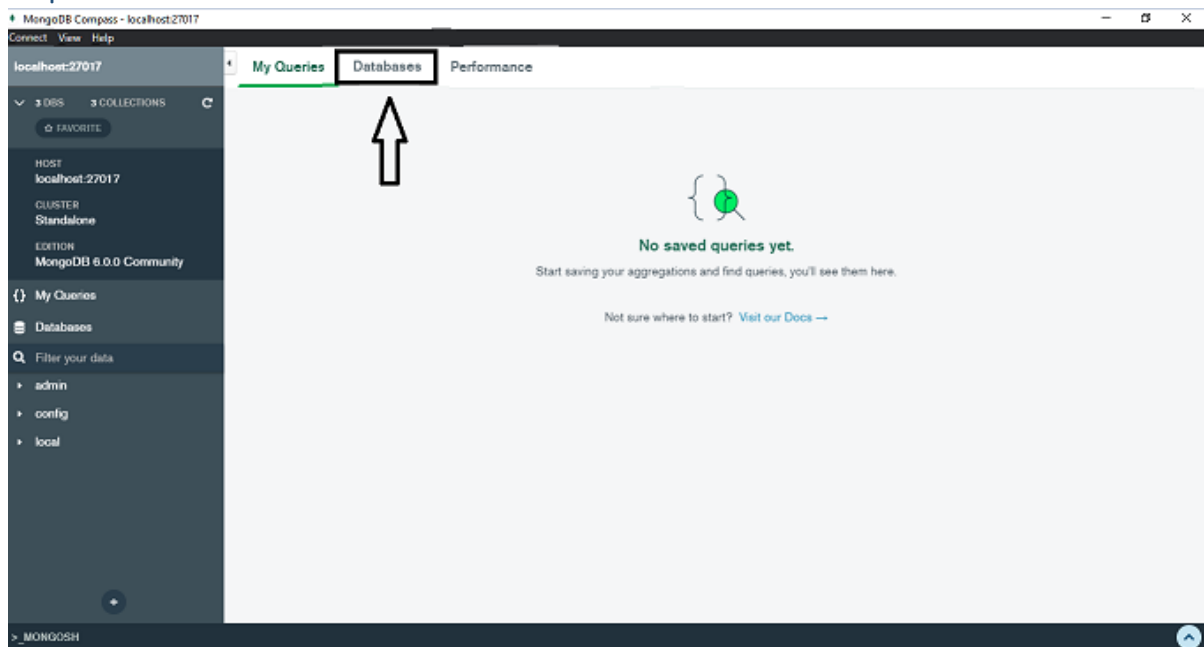
Showing existing Databases

```
show databases
```

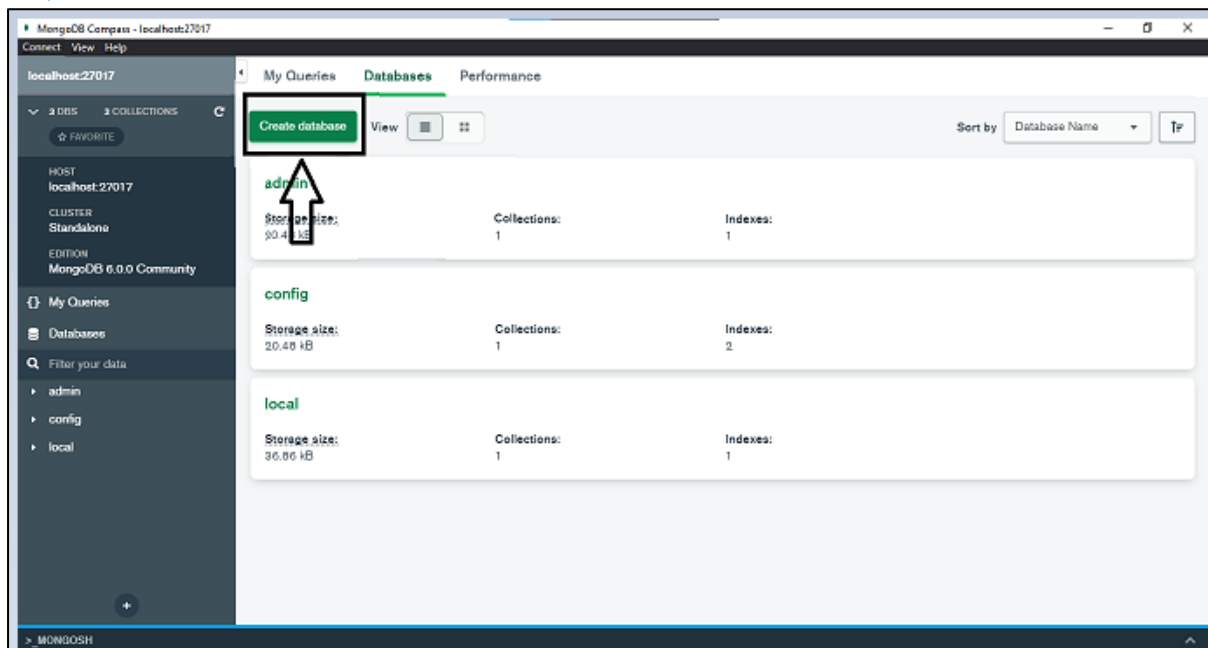
```
> _MONGOSH  
  
> show databases  
< admin    40.00 KiB  
  config   60.00 KiB  
  local    72.00 KiB  
  
test>
```

Creating New Database

Step 1: Click on the Databases Tab



Step 2: Click on Create Database



Step 3: Assign the database name and the collection name and click on the Create Database

Create Database

1. Give the Database name (e.g MSC)

Database Name

Collection Name

> Advanced Collection Options (e.g. Time-Series, Capped, Clustered collections)

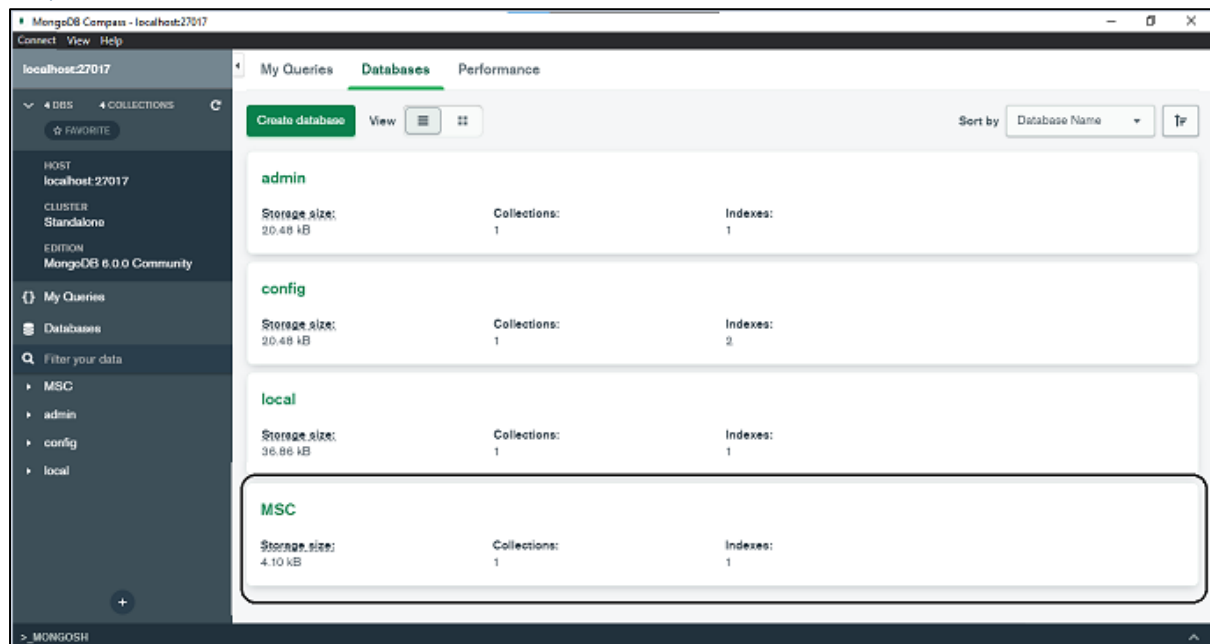
Before MongoDB can save your new database, a collection name must also be specified at the time of creation. [More Information](#)

2. Give the Collection Name (e.g STD_DATA)

3. Click on Create Database

Cancel Create Database

Step 4: After the Create Database you will be able to see the database



Switching database to MSC

```
use MSC
```

```
> _MONGOSH
> use MSC
< 'switched to db MSC'
MSC>
```

Q2. Create a new collection Students and add values and display all records.

→

Inserting Records into the collection STD_DATA

```
db.STD_DATA.insertMany([{"Name": "Harsh", Age: 22, Hobby: ["Reading", "Cricket"]}, {"Name": "Ashok", Age: 23, Hobby: ["Singing", "Cricket"]}, {"Name": "ASHA", Age: 21, Hobby: ["Singing"]}, {"Name": "John", Age: 22, Hobby: ["Reading", "Painting"]}])
```

```
>_MONGOSH
> use MSC
< 'switched to db MSC'
> db.STD_DATA.insertMany([({Name: "Harsh", Age: 22, Hobby: ["Reading","Cricket"]},{Name: "Ashok", Age: 23, Hobby: ["Singing","Cricket"]},{Name: "ASHA", Age: 21, Hobby: ["Sin
< { acknowledged: true,
  insertedIds:
    { '0': ObjectId("62df6bb1962590bd333bfe66"),
      '1': ObjectId("62df6bb1962590bd333bfe67"),
      '2': ObjectId("62df6bb1962590bd333bfe68"),
      '3': ObjectId("62df6bb1962590bd333bfe69") } }
MSC>
```

Displaying all the records

```
db.STD_DATA.find()
```

```
> db.STD_DATA.find()
< { _id: ObjectId("62df6bb1962590bd333bfe66"),
  Name: 'Harsh',
  Age: 22,
  Hobby: [ 'Reading', 'Cricket' ] }
{ _id: ObjectId("62df6bb1962590bd333bfe67"),
  Name: 'Ashok',
  Age: 23,
  Hobby: [ 'Singing', 'Cricket' ] }
{ _id: ObjectId("62df6bb1962590bd333bfe68"),
  Name: 'ASHA',
  Age: 21,
  Hobby: [ 'Singing' ] }
{ _id: ObjectId("62df6bb1962590bd333bfe69"),
  Name: 'John',
  Age: 22,
  Hobby: [ 'Reading', 'Painting' ] }
MSC>
```

Q3. Display details of Ashok

→

```
db.STD_DATA.find({Name:"Ashok"})
```

```
> _MONGOSH
> db.STD_DATA.find({Name:"Ashok"})
< { _id: ObjectId("62df6bb1962590bd333bfe67"),
  Name: 'Ashok',
  Age: 23,
  Hobby: [ 'Singing', 'Cricket' ] }
MSC>
```

Q4. Update age of John to 20 keep rest of the data same

→

Updating the record of John

```
db.STD_DATA.findOneAndUpdate({Name:"John"},{$set :{Age:20}})
```

```
> _MONGOSH
> db.STD_DATA.findOneAndUpdate({Name:"John"},{$set :{Age:20}})
< { _id: ObjectId("62df6e42962590bd333bfe6d"),
  Name: 'John',
  Age: 22,
  Hobby: [ 'Reading', 'Painting' ] }
MSC>
```

Displaying the updated record for John

```
db.STD_DATA.find({Name:"John"})
```

```
> _MONGOSH
> db.STD_DATA.find({Name:"John"})
< { _id: ObjectId("62df6e42962590bd333bfe6d"),
  Name: 'John',
  Age: 20,
  Hobby: [ 'Reading', 'Painting' ] }
MSC>
```

Q5. Update hobby of Harsh as Dancing instead of Reading.

→

Updating the record of Harsh

```
db.STD_DATA.findOneAndUpdate({"Name":"Harsh"}, { $set:
{"Hobby.$[element]":"Dancing"}},{ arrayFilters: [{ element: "Reading" }]])
```

```
> _MONGOSH
> db.STD_DATA.findOneAndUpdate({"Name":"Harsh"}, { $set: {"Hobby.$[element]":"Dancing"}},{ arrayFilters: [{ element: "Reading" }]])
< { _id: ObjectId("62df6e42962590bd333bfe6a"),
  Name: 'Harsh',
  Age: 22,
  Hobby: [ 'Reading', 'Cricket' ] }
MSC>
```

Displaying Records of Harsh

```
> _MONGOSH
> db.STD_DATA.find({Name:"Harsh"})
< { _id: ObjectId("62df6e42962590bd333bfe6a"),
  Name: 'Harsh',
  Age: 22,
  Hobby: [ 'Dancing', 'Cricket' ] }
MSC>
```

Q6. Display name whose age is 22.

→

```
db.STD_DATA.find({Age:20},{Name:1})
```



```
> _MONGOSH  
  
> db.STD_DATA.find({Age:20},{Name:1})  
< { _id: ObjectId("62df6e42962590bd333bfe6d"), Name: 'John' }  
MSC>
```

Q7. Delete record of John.

→

```
db.STD_DATA.deleteOne({Name:"John"})
```

```
> _MONGOSH  
  
> db.STD_DATA.deleteOne({Name:"John"})  
< { acknowledged: true, deletedCount: 1 }  
MSC>
```

Q8. Update Age of Ashok first occurrence as 19.

→

Updating Record

```
db.STD_DATA.findOneAndUpdate({Name:"Ashok"},{$set :{Age:19}})  
  
db.STD_DATA.find({Name:"Ashok"})
```

```
> _MONGOSH
> db.STD_DATA.findOneAndUpdate({Name:"Ashok"},{$set :{Age:19}})
< { _id: ObjectId("62df6e42962590bd333bfe6b"),
  Name: 'Ashok',
  Age: 23,
  Hobby: [ 'Singing', 'Cricket' ] }
> db.STD_DATA.find({Name:"Ashok"})
< { _id: ObjectId("62df6e42962590bd333bfe6b"),
  Name: 'Ashok',
  Age: 19,
  Hobby: [ 'Singing', 'Cricket' ] }
{ _id: ObjectId("62df758a962590bd333bfe6e"),
  Name: 'Ashok',
  Age: 20,
  Hobby: [ 'Singing', 'Cricket' ] }
MSC>
```

Q9. Update Age of Ashok, all Occurrences as 15.

→

```
db.STD_DATA.updateMany({Name:"Ashok"},{$set :{Age:15}})

db.STD_DATA.find({Name:"Ashok"})
```

```
> _MONGOSH
> db.STD_DATA.updateMany({Name:"Ashok"},{$set :{Age:15}})
< { acknowledged: true,
  insertedId: null,
  matchedCount: 2,
  modifiedCount: 2,
  upsertedCount: 0 }
> db.STD_DATA.find({Name:"Ashok"})
< { _id: ObjectId("62df6e42962590bd333bfe6b"),
  Name: 'Ashok',
  Age: 15,
  Hobby: [ 'Singing', 'Cricket' ] }
{ _id: ObjectId("62df758a962590bd333bfe6e"),
  Name: 'Ashok',
  Age: 15,
  Hobby: [ 'Singing', 'Cricket' ] }
MSC>
```

Q10. Add mobile no. of Harsh.

→

```
db.STD_DATA.updateMany({Name:"Harsh"},{$set :{Mobile:9372685907}})

db.STD_DATA.find({Name:"Harsh"})
```

```
> _MONGOSH

> db.STD_DATA.updateMany({Name:"Harsh"},{$set :{Mobile:9372685907}})
< { acknowledged: true,
    insertedId: null,
    matchedCount: 1,
    modifiedCount: 1,
    upsertedCount: 0 }
> db.STD_DATA.find({Name:"Harsh"})
< { _id: ObjectId("62df6e42962590bd333bfe6a"),
    Name: 'Harsh',
    Age: 22,
    Hobby: [ 'Dancing', 'Cricket' ],
    Mobile: 9372685907 }
MSC>
```

Q11. Display Record whose age is 22 and hobby as Dancing.

→

```
db.STD_DATA.find({$and:[{Age:22},{Hobby: "Dancing"}]}))
```

```
<
> db.STD_DATA.find({$and:[{Age:22},{Hobby: "Dancing"}]})
< { _id: ObjectId("62df6e42962590bd333bfe6a"),
    Name: 'Harsh',
    Age: 22,
    Hobby: [ 'Dancing', 'Cricket' ],
    Mobile: 9372685907 }
MSC>
```

Q12. Display Record whose age is 15 or hobby as Singing

→

```
db.STD_DATA.find({$or: [{Age:15},{ Hobby:["Singing"]}]}))
```

```
> _MONGOSH
> db.STD_DATA.find({$or: [{Age:15},{ Hobby:["Singing"]}]}))
< { _id: ObjectId("62df6e42962590bd333bfe6b"),
  Name: 'Ashok',
  Age: 15,
  Hobby: [ 'Singing', 'Cricket' ] }
{ _id: ObjectId("62df6e42962590bd333bfe6c"),
  Name: 'ASHA',
  Age: 21,
  Hobby: [ 'Singing' ] }
{ _id: ObjectId("62df758a962590bd333bfe6e"),
  Name: 'Ashok',
  Age: 15,
  Hobby: [ 'Singing', 'Cricket' ] }
MSC>
```

Q13. Display records from the collection by skipping first 2 records.

→

```
db.STD_DATA.find({}).skip(2).limit(2)
```

```
> _MONGOSH
> db.STD_DATA.find({}).skip(2).limit(2)
< { _id: ObjectId("62df6e42962590bd333bfe6c"),
  Name: 'ASHA',
  Age: 21,
  Hobby: [ 'Singing' ] }
{ _id: ObjectId("62df758a962590bd333bfe6e"),
  Name: 'Ashok',
  Age: 15,
  Hobby: [ 'Singing', 'Cricket' ] }
MSC>
```

Q14. Display records by sorting names.

→

```
db.STD_DATA.find({}).sort({name:-1})
```

```
> _MONGOSH
> db.STD_DATA.find({}).sort({name:-1})
< { _id: ObjectId("62df6e42962590bd333bfe6a"),
  Name: 'Harsh',
  Age: 22,
  Hobby: [ 'Dancing', 'Cricket' ],
  Mobile: 9372685907 }
{ _id: ObjectId("62df6e42962590bd333bfe6b"),
  Name: 'Ashok',
  Age: 15,
  Hobby: [ 'Singing', 'Cricket' ] }
{ _id: ObjectId("62df6e42962590bd333bfe6c"),
  Name: 'ASHA',
  Age: 21,
  Hobby: [ 'Singing' ] }
{ _id: ObjectId("62df758a962590bd333bfe6e"),
  Name: 'Ashok',
  Age: 15,
  Hobby: [ 'Singing', 'Cricket' ] }
MSC> |
```

Q15. Count the number of records into the collection

→

```
db.STD_DATA.find({}).count({})
```

```
> _MONGOSH
> db.STD_DATA.find({}).count({})
< 4
MSC> |
```

Q16) Show existing Databases and create a new DB MSC.



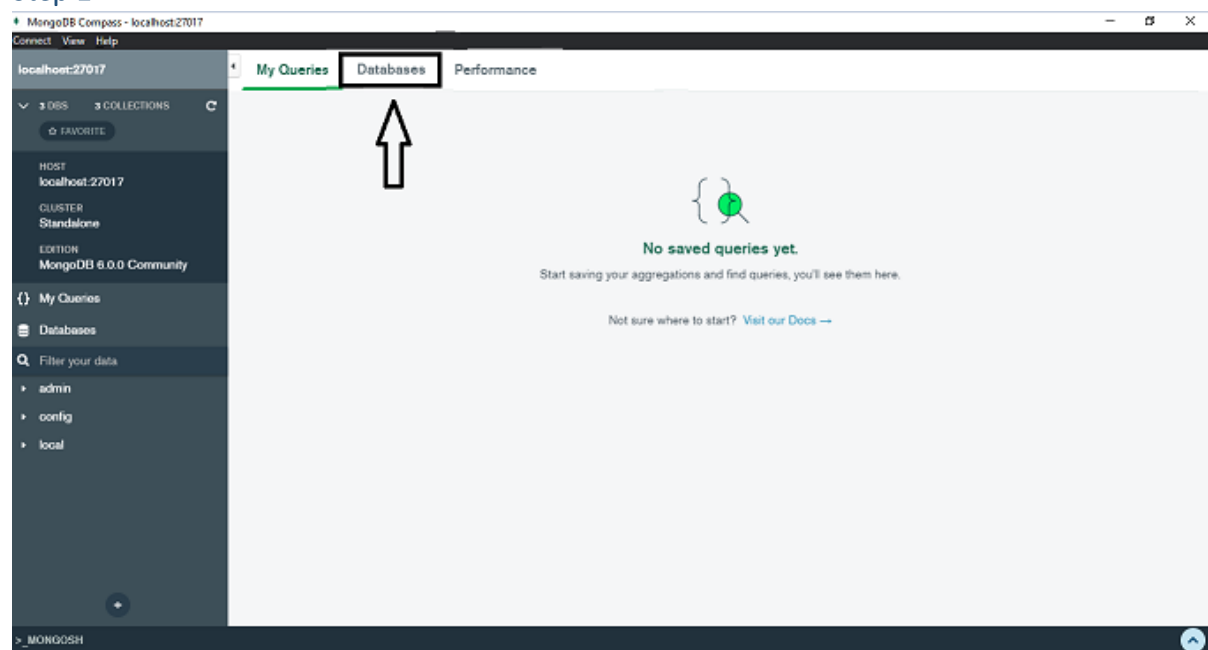
Showing existing Databases

```
show databases
```

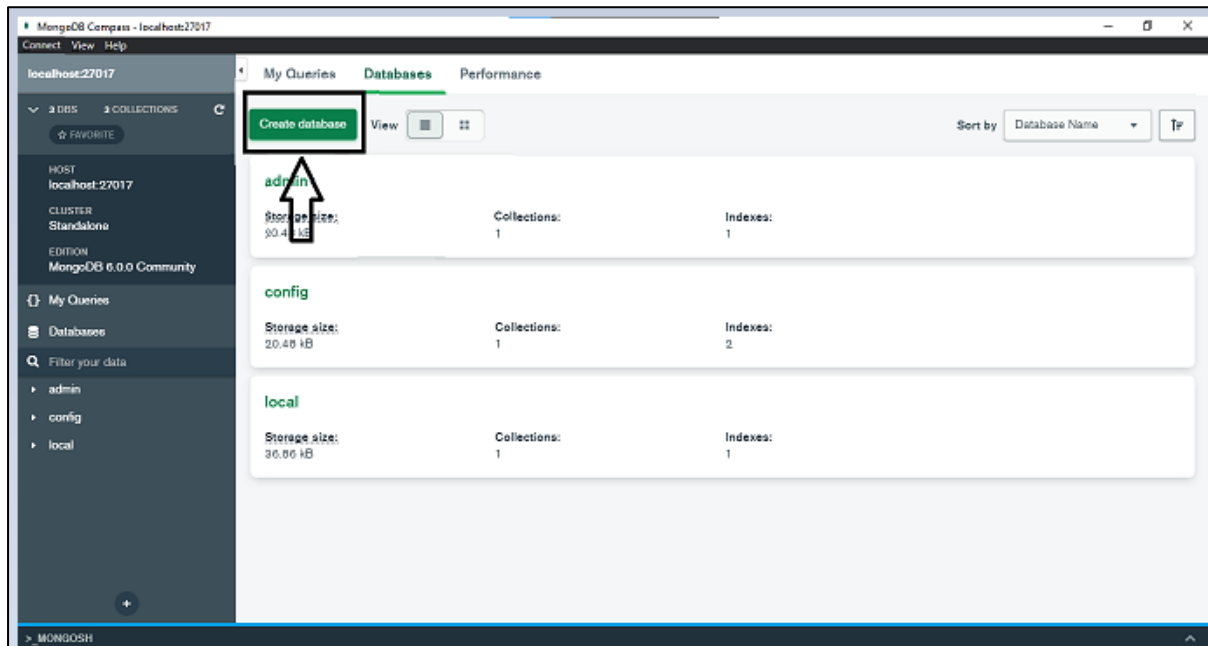
```
>_MONGOSH  
  
> show databases  
< admin    40.00 KiB  
   config   60.00 KiB  
   local    72.00 KiB  
  
test>
```

Creating New Database

Step 1: Click on the Databases Tab



Step 2: Click on Create Database



Step 3: Assign the database name and the collection name and click on the Create Database

Create Database 1. Give the database name (e.g: MSC)

Database Name

Collection Name

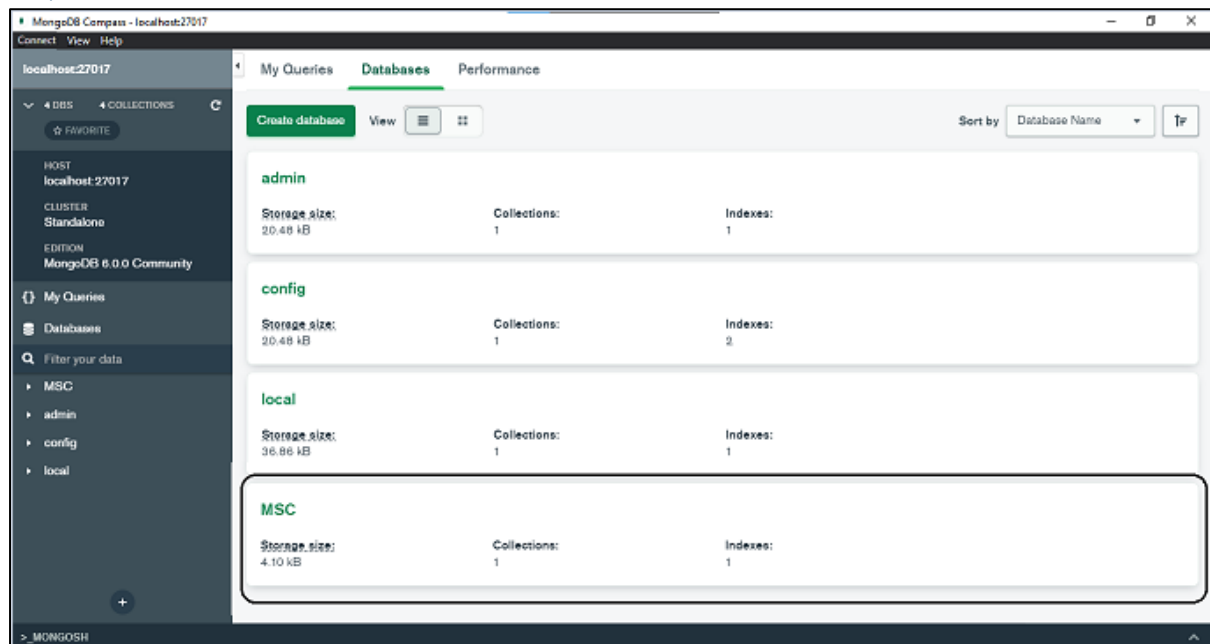
> Advanced Collection Options (e.g. Time-Series, Capped, Clustered collections)

Before MongoDB can save your new database, a collection name must also be specified at the time of creation. [More Information](#)

2. Give the Collection Name (e.g STD-Marks_Details)

3. Click on Create Database

Step 4: After the Create Database you will be able to see the database



Switching database to MSC

```
use MSC
```

```
> _MONGOSH
> use MSC
< 'switched to db MSC'
MSC>
```

Q17) Create a MongoDB containing marks of students for subjects like English, Maths and Computer.

Insert 3 documents where

- 1) First document is having marks of Maths and Computer
- 2) Second document is having marks of Maths and English
- 3) Third document is having marks of all subjects

→

```
db.STD_MARKS_DETAILS.insertMany([{'Name':'John','Roll_No':01,'Subject':{'Maths':91,'Computer':85}},{'Name':'Tom','Roll_No':02,'Subject':{'Maths':83,'English':45}},{'Name':'Bob','Roll_No':03,'Subject':{'Maths':76,'English':70,'Computer':85}}])
```



```
> _MONGOSH
> db.STD_MARKS_DETAILS.insertMany([({Name:"John",Roll_No:01,Subject:{'Maths':91,'Computer':85}},{Name:"Tom",Roll_No:02,Subject:{'Maths':83,'English':45}},{Name:"Bob",Roll_No:03,Subject:{'Maths':88,'English':92}}])
< { acknowledged: true,
  insertedIds:
    { '0': ObjectId("62e0b55705c69448b5713918"),
      '1': ObjectId("62e0b55705c69448b5713919"),
      '2': ObjectId("62e0b55705c69448b571391a") } }
MSC>
```

Q18) Update marks of 1st document English as 85.

→

```
db.STD_MARKS_DETAILS.updateOne({Name:'John'},{$set:{'Subject.English':85}})

db.STD_MARKS_DETAILS.find({Name:'John'})
```

```
> _MONGOSH

> db.STD_MARKS_DETAILS.updateOne({Name:'John'},{$set:{'Subject.English':85}})
< { acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0 }
> db.STD_MARKS_DETAILS.find({Name:'John'})
< { _id: ObjectId("62e0b55705c69448b5713918"),
  Name: 'John',
  Roll_No: 1,
  Subject: { Maths: 91, Computer: 85, English: 85 } }
MSC>
```

Q19) Retrieve does contain marks of English as 85 and Maths as 91.

→

```
db.STD_MARKS_DETAILS.find({$and:[{'Subject.Maths':91},{'Subject.English':85}]})
```

```
> _MONGOSH  
  
> db.STD_MARKS_DETAILS.find({$and:[{'Subject.Maths':91},{'Subject.English':85}]})  
< { _id: ObjectId("62e0b55705c69448b5713918"),  
  Name: 'John',  
  Roll_No: 1,  
  Subject: { Maths: 91, Computer: 85, English: 85 } }  
MSC> |
```

Practical 2

Aim: Installation of Hadoop and java in windows.



Installation on java jdk and jre

Step 1: Download SE Development Kit from the given link

<https://www.oracle.com/in/java/technologies/javase/javase8-archive-downloads.html>

Operating System	Size	Download Link
Solaris SPARC 64-bit (SVR4 package)	125.09 MB	jdk-8u202-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	88.1 MB	jdk-8u202-solaris-sparcv9.tar.gz
Solaris x64 (SVR4 package)	124.37 MB	jdk-8u202-solaris-x64.tar.Z
Solaris x64	85.38 MB	jdk-8u202-solaris-x64.tar.gz
Windows x86	201.64 MB	jdk-8u202-windows-i586.exe
Windows x64	211.58 MB	jdk-8u202-windows-x64.exe

Java SE Runtime Environment 8u202
This software is licensed under the Oracle Binary Code License Agreement for Java SE Platform Products

Click on Download

Step 2: Download Java SE Runtime Environment 8u202 from the given link

<https://www.oracle.com/in/java/technologies/javase/javase8-archive-downloads.html>

Operating System	Size	Download Link
Mac OS X x64	69.37 MB	jre-8u202-macosx-x64.tar.gz
Solaris SPARC 64-bit	46.07 MB	jre-8u202-solaris-sparcv9.tar.gz
Solaris x64	43.36 MB	jre-8u202-solaris-x64.tar.gz
Windows x86 Online	1.83 MB	jre-8u202-windows-i586-iftw.exe
Windows x86 Offline	65.73 MB	jre-8u202-windows-i586.exe
Windows x86	68.4 MB	jre-8u202-windows-i586.tar.gz
Windows x64	73.7 MB	jre-8u202-windows-x64.exe
Windows x64	73.25 MB	jre-8u202-windows-x64.tar.gz

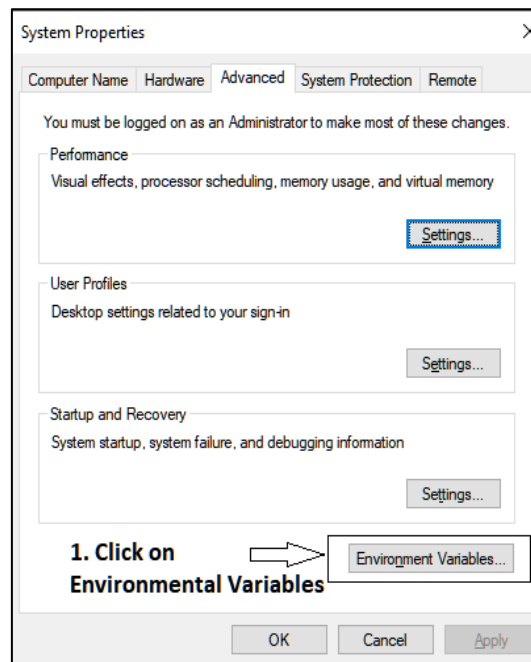
Server JRE (Java SE Runtime Environment) 8u202

Click on Download

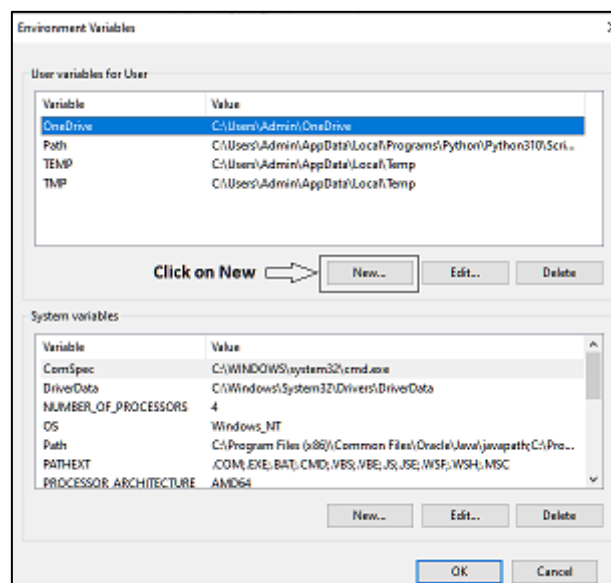
Setting up the JDK in Environmental Variable

Step 1: Start → Edit the system environment variables

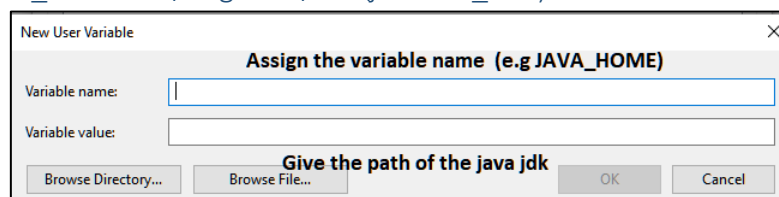
Step 2: Click on the environment Variables

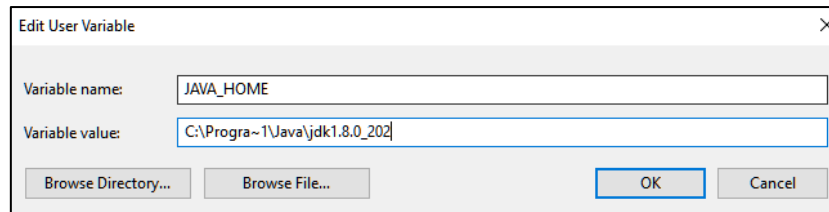


Step 3: Create the new user variables

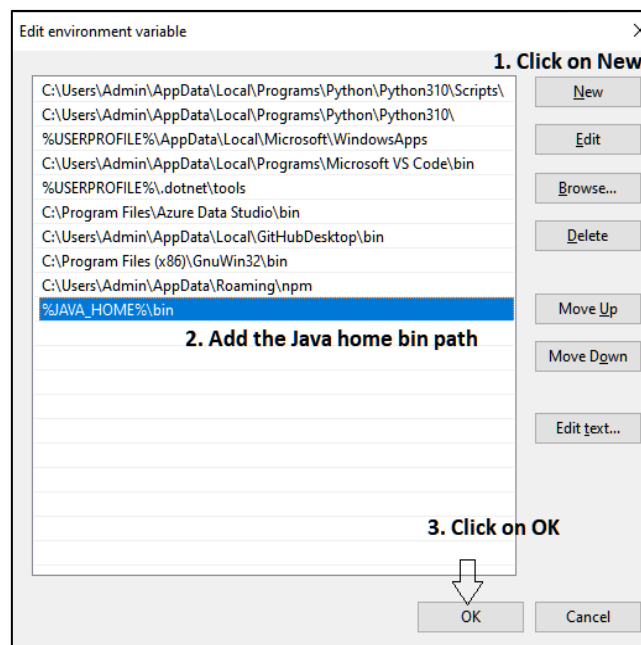
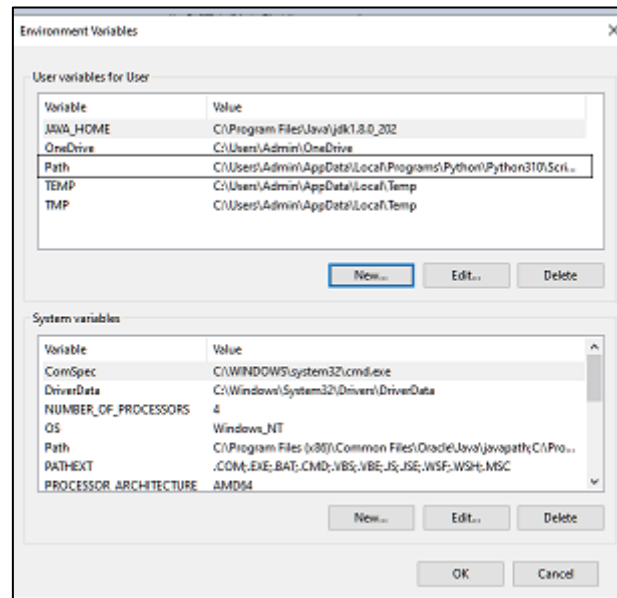


Step 4: Give the Variable Name and java jdk path (NOTE: Rename the path C:\Program files\Java\jdk1.8.0_202 to C:\Progra~1\Java\jdk1.8.0_202)

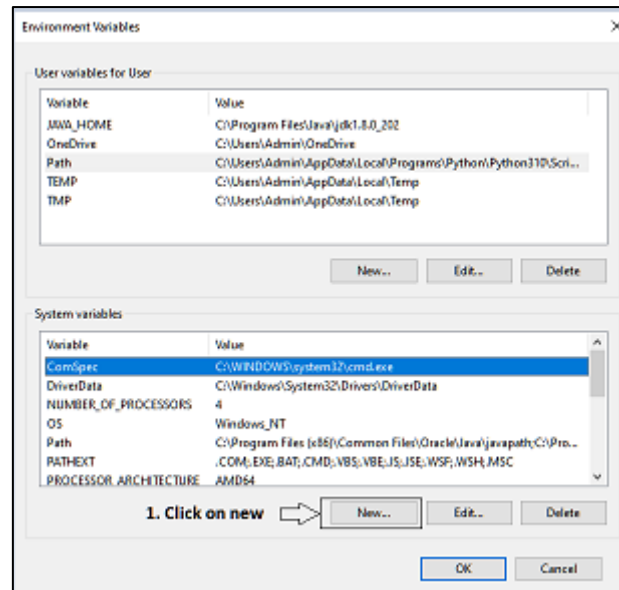




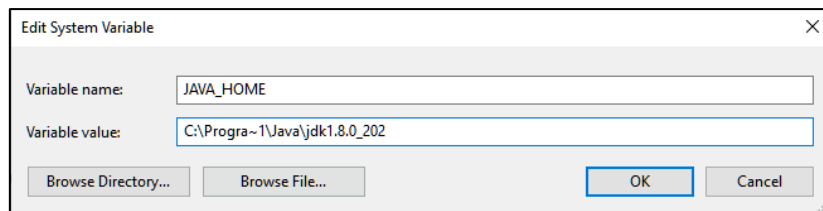
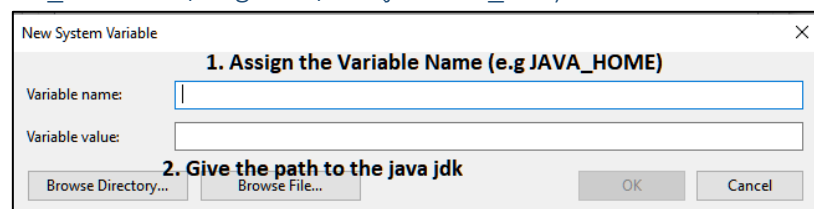
Step 5: Add the Variable into the Path



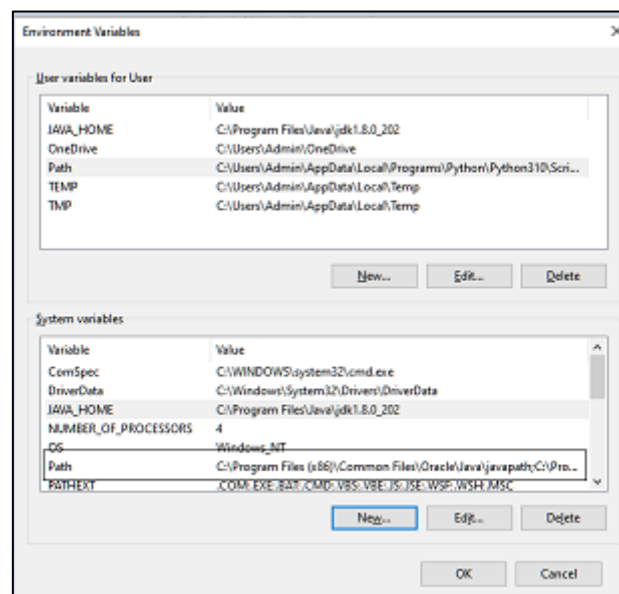
Step 6: Create the new system variables

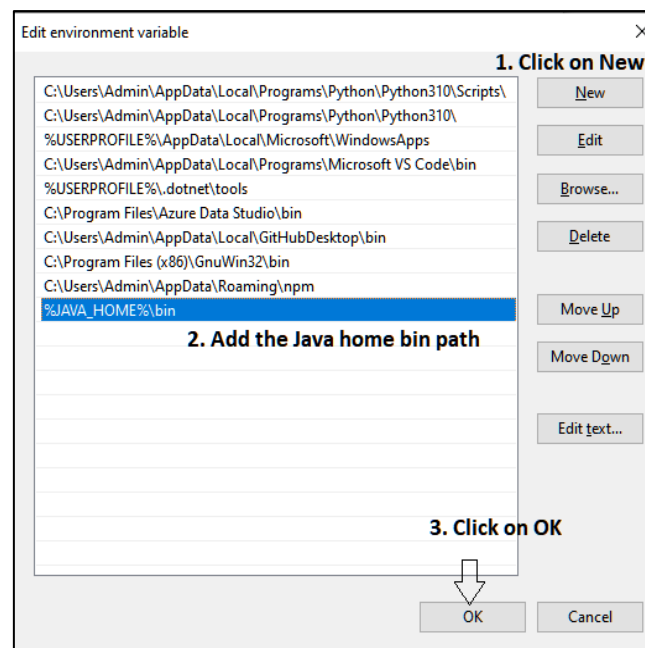


Step 7: Give the Variable Name and java jdk path (NOTE: Rename the path C:\Program files\Java\jdk1.8.0_202 to C:\Progra~1\Java\jdk1.8.0_202)



Step 8: Add the Variable into the Path





Installation of Hadoop

Step 1: Download Apache Hadoop from the given link and click on the binary
<https://hadoop.apache.org/releases.html>

Apache Hadoop Download Documentation Community Development Help Apache Software Foundation

Download

Hadoop is released as source code tarballs with corresponding binary tarballs for convenience. The downloads are distributed via mirror sites and should be checked for tampering using GPG or SHA-512.

Version	Release date	Source download	Click to Download		Release notes
			Binary download		
3.2.4	2022 Jul 22	source (checksum signature)	binary (checksum signature)		Announcement
2.10.2	2022 May 31	source (checksum signature)	binary (checksum signature)		Announcement
3.3.3	2022 May 17	source (checksum signature)	binary (checksum signature) binary-aarch64 (checksum signature)		Announcement

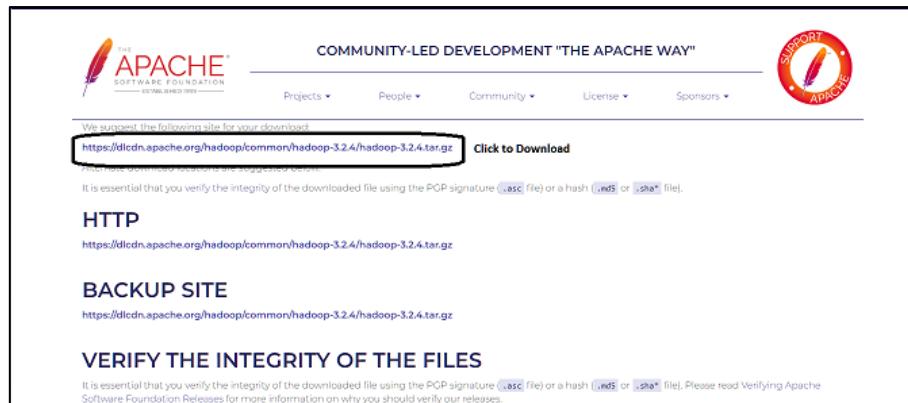
To verify Hadoop releases using GPG:

1. Download the release hadoop-X.Y.Z-src.tar.gz from a mirror site.
2. Download the signature file hadoop-X.Y.Z-src.tar.gz.asc from Apache.
3. Download the Hadoop KEYS file.
4. gpg --import KEYS
5. gpg --verify hadoop-X.Y.Z-src.tar.gz.asc

To perform a quick check using SHA-512:

1. Download the release hadoop-X.Y.Z-src.tar.gz from a mirror site.
2. Download the checksum hadoop-X.Y.Z-src.tar.gz.sha512 or hadoop-X.Y.Z-src.tar.gz.md5 from Apache.
3. shasum -a 512 hadoop-X.Y.Z-src.tar.gz

Step 2: Click the link to download the zip

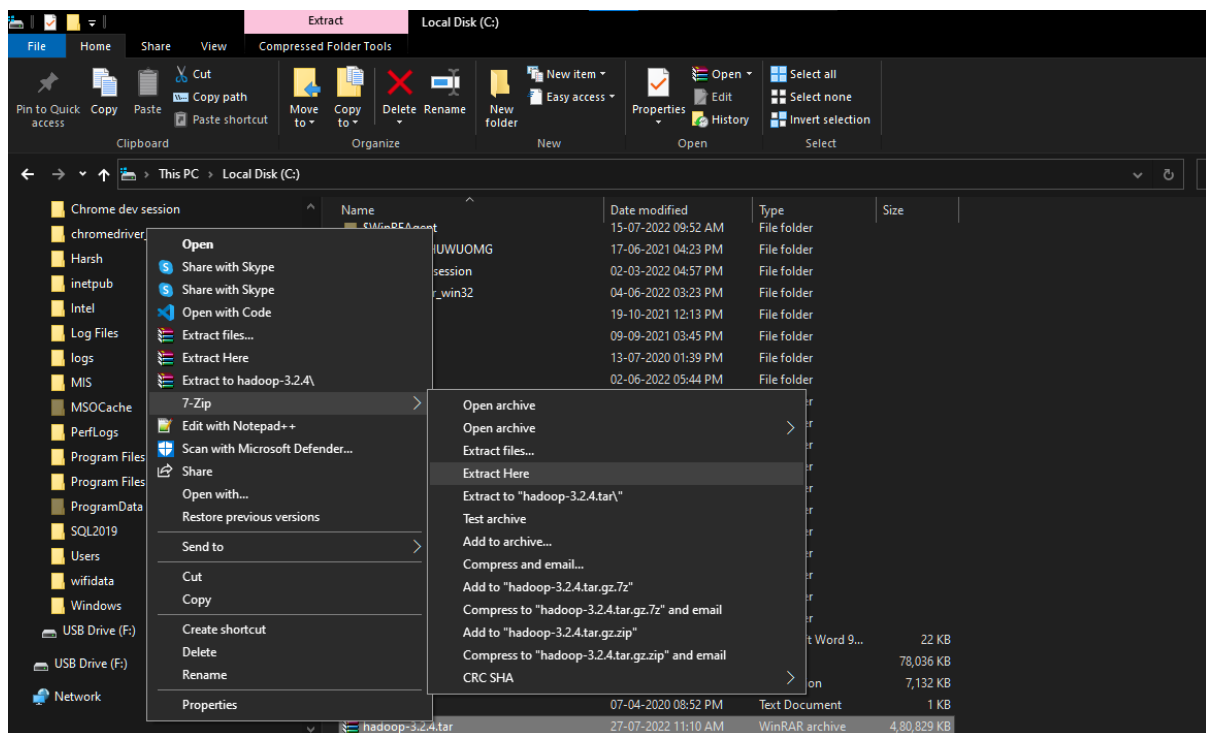


Installation of the 7zip

Step 1: Download the 7zip and install into the system <https://www.7-zip.org/download.html>

Extracting Hadoop zip using 7Zip

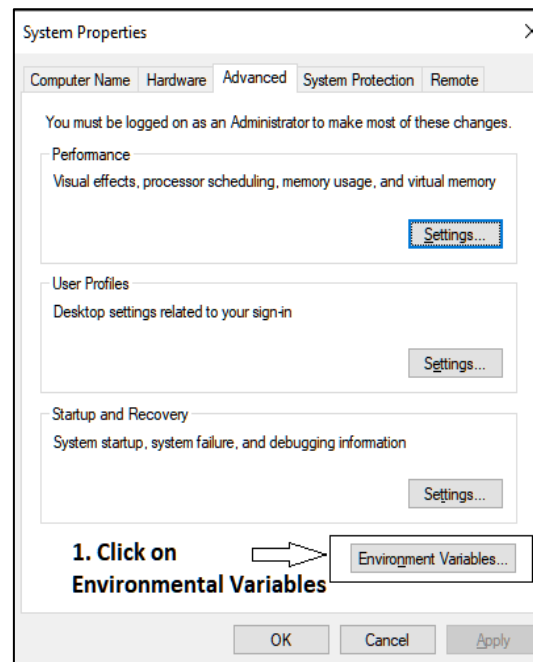
Step 1: Right Click on the zip → Click on 7-zip → Extract Here



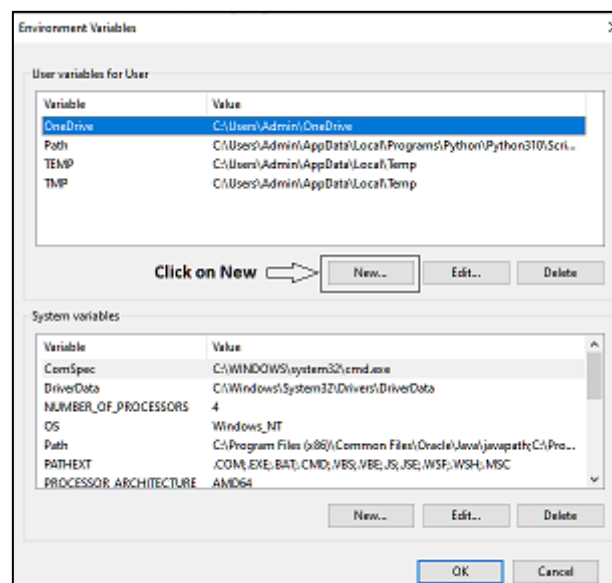
Setting up the Hadoop in Environmental Variable

Step 1: Start → Edit the system environment variables

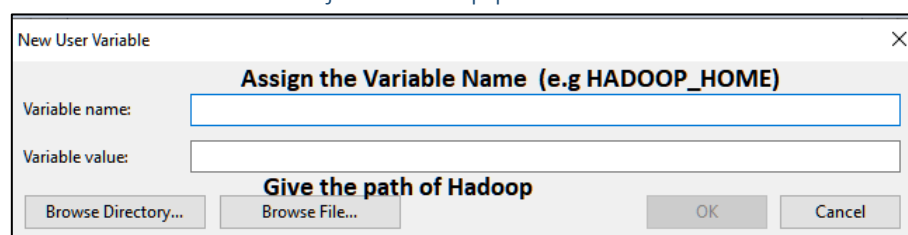
Step 2: Click on the environment Variables

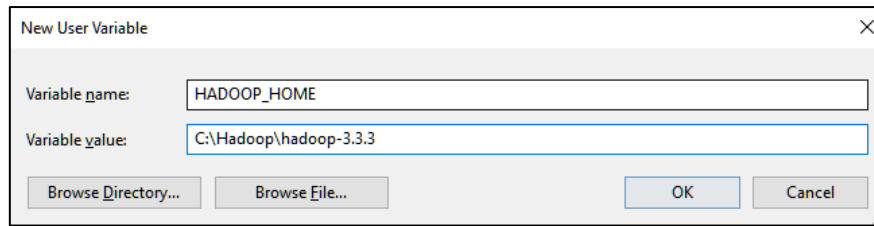


Step 3: Create the new user variables

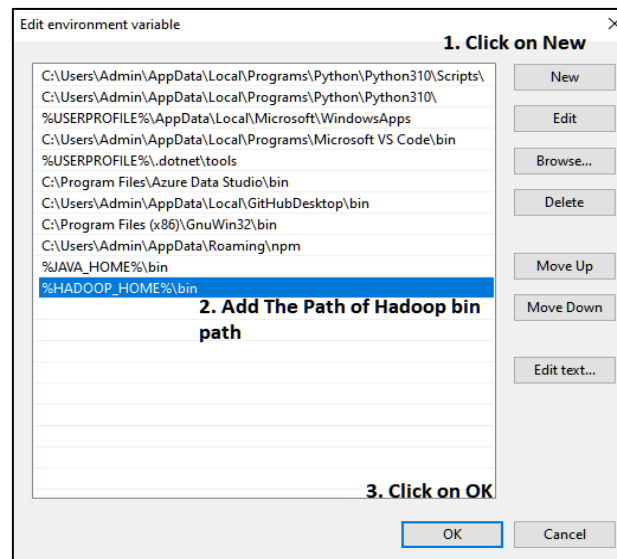
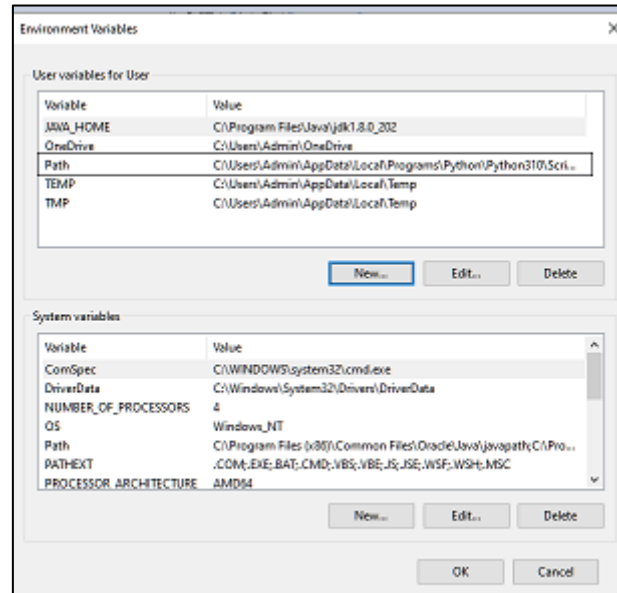


Step 4: Give the Variable Name and java Hadoop path

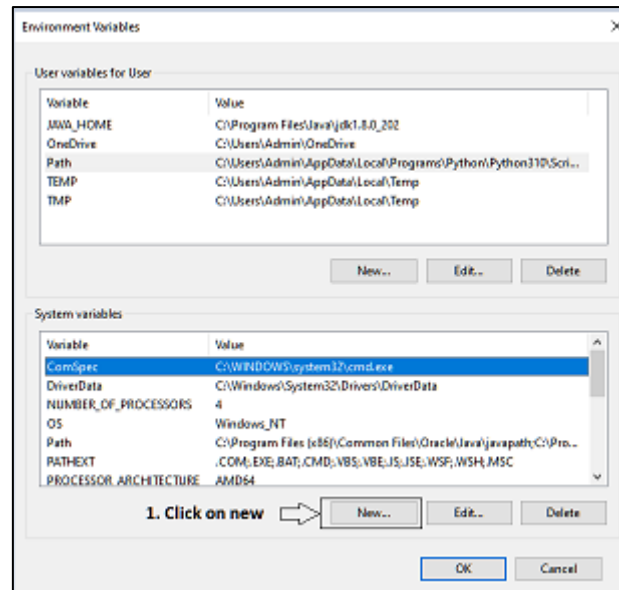




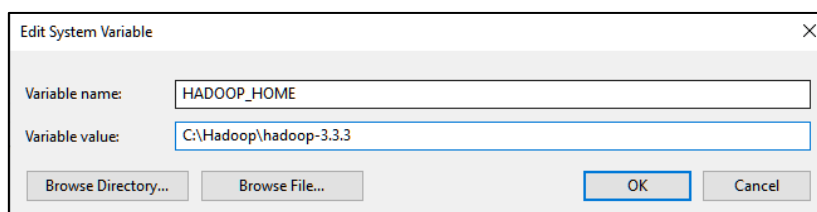
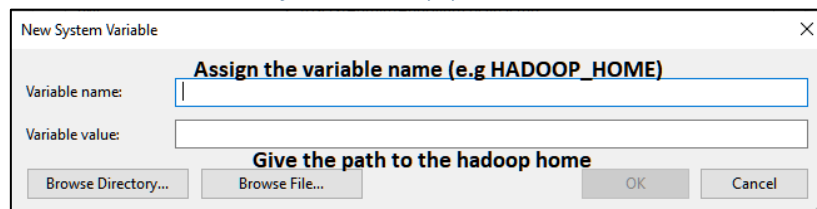
Step 5: Add the Variable into the Path



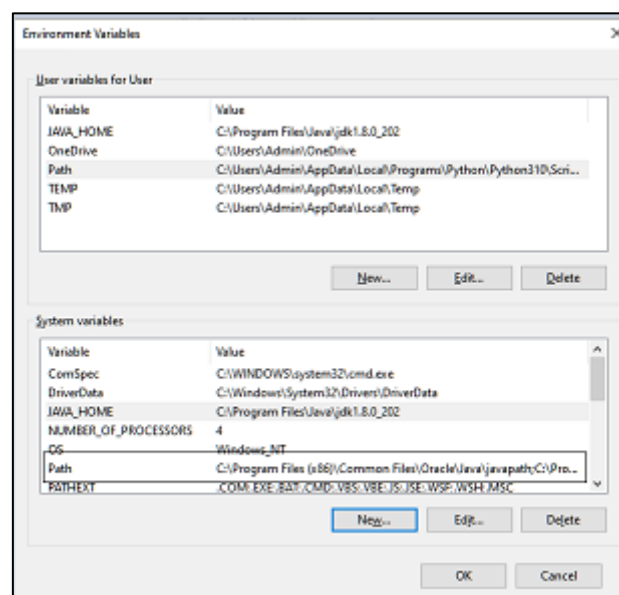
Step 6: Create the new system variables

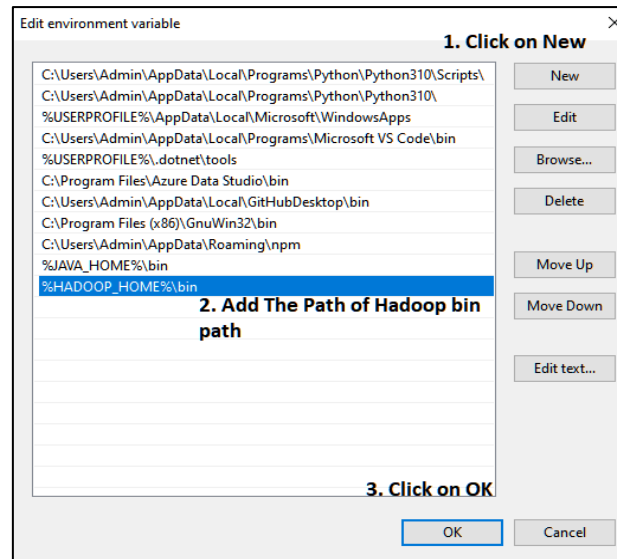


Step 7: Give the Variable Name and java Hadoop path



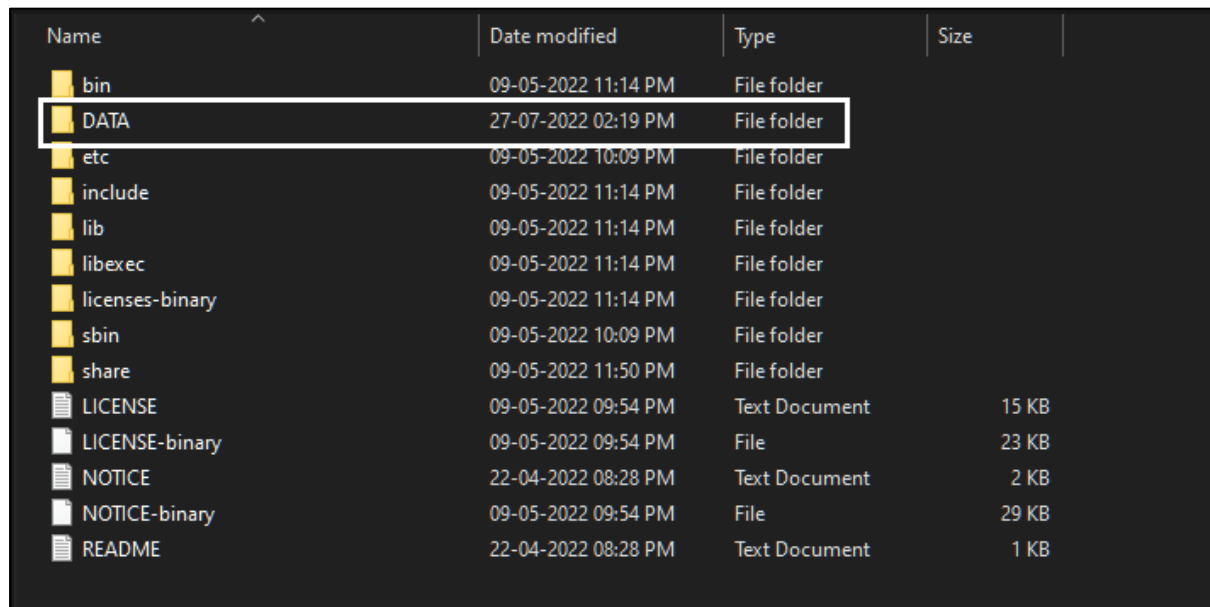
Step 8: Add the Variable into the Path



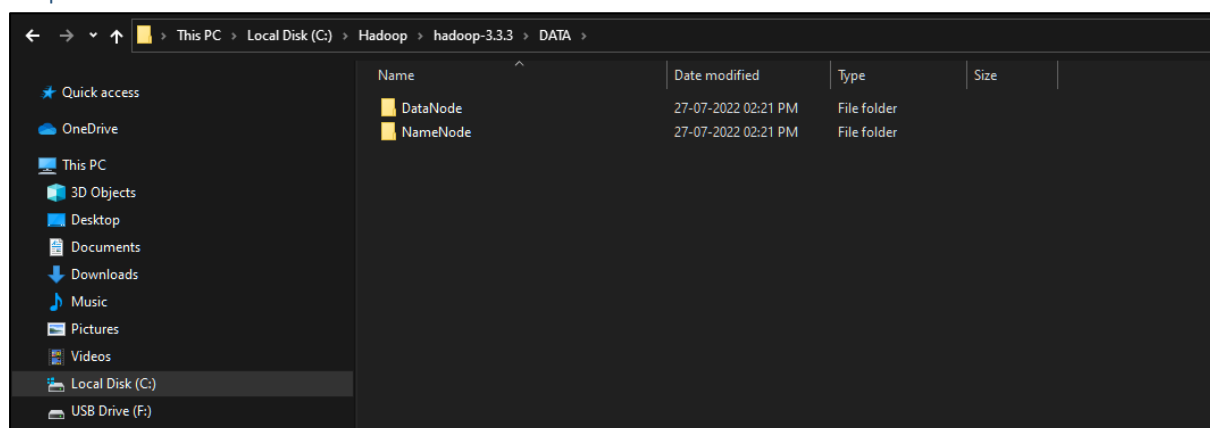


Setting up Hadoop

Step 1: Create the new Folder with the name DATA



Step 2: Go inside the data folder and create 2 folder i.e. NameNode and DataNode



Step 3: open the file `hdfs-site.xml` `C:\Hadoop\hadoop-3.3.3\etc\hadoop`

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
  <property>
    <name>dfs.replication</name>
    <value>3</value>
  </property>
  <property>
<name>dfs.namenode.name.dir</name>
<value>C:\Hadoop\hadoop-3.3.3\DATA\NameNode</value>
</property>

<property>
<name>dfs.datanode.data.dir</name>
<value>C:\Hadoop\hadoop-3.3.3\DATA\DataNode</value>
</property>

<property>
  <name>dfs.permissions</name>
  <value>>false</value>
</property>
</configuration>
```

```
</> hdfs-site.xml X
C:\Hadoop\hadoop-3.3.3\etc\hadoop\hdfs-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3
4
5 <configuration>
6   <property>
7     <name>dfs.replication</name>
8     <value>3</value>
9   </property>
10  <property>
11    <name>dfs.namenode.name.dir</name>
12    <value>C:\Hadoop\hadoop-3.3.3\DATA\NameNode</value>
13  </property>
14
15  <property>
16    <name>dfs.datanode.data.dir</name>
17    <value>C:\Hadoop\hadoop-3.3.3\DATA\DataNode</value>
18  </property>
19
20  <property>
21    <name>dfs.permissions</name>
22    <value>false</value>
23  </property>
24 </configuration>
```

Step 4: open the file `core-site.xml` C:\Hadoop\hadoop-3.3.3\etc\hadoop

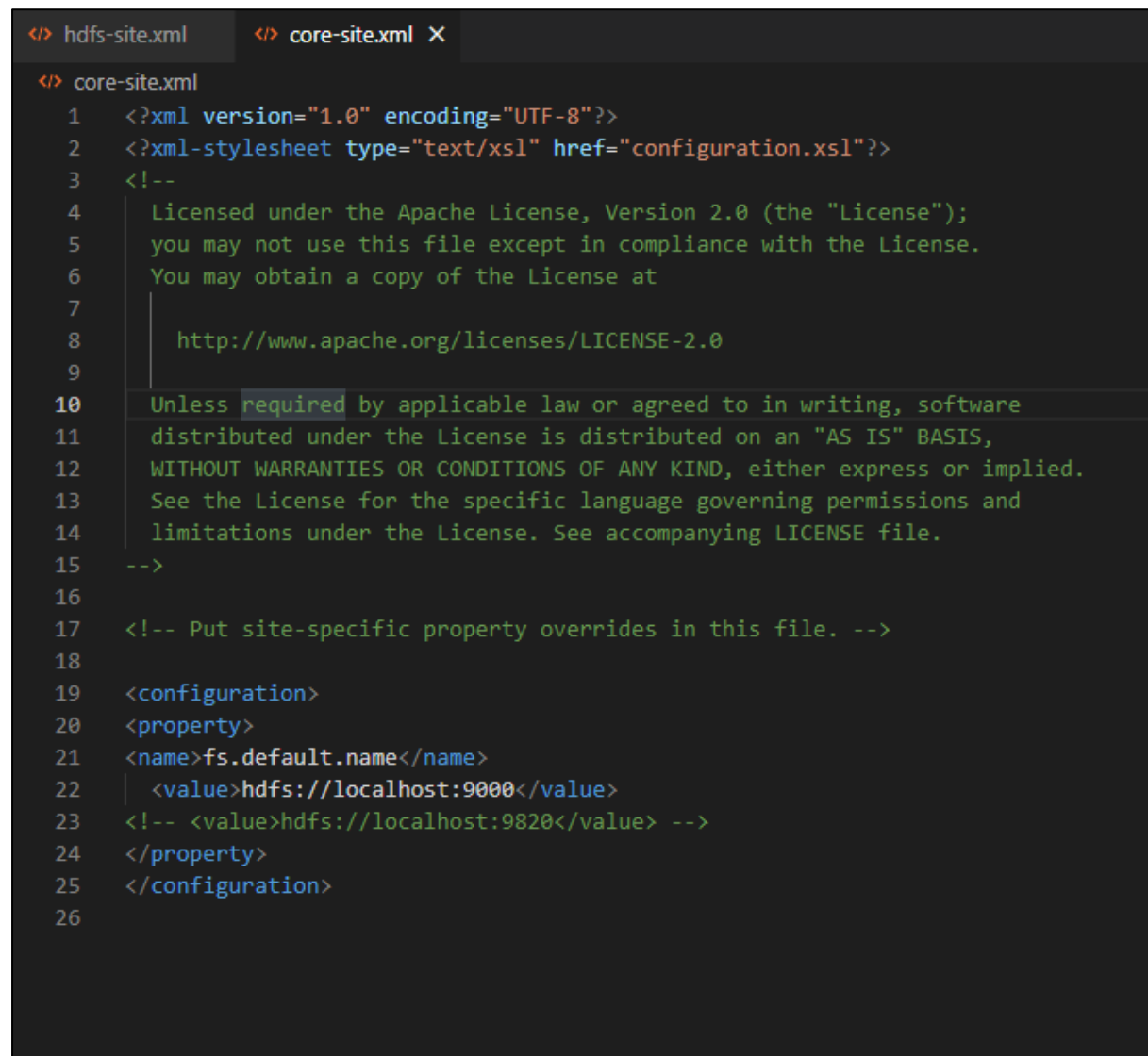
```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
```

```
<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
<name>fs.default.name</name>
  <value>hdfs://localhost:9000</value>
<!-- <value>hdfs://localhost:9820</value> -->
</property>
</configuration>
```



```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
<name>fs.default.name</name>
  <value>hdfs://localhost:9000</value>
<!-- <value>hdfs://localhost:9820</value> -->
</property>
</configuration>
```

Step 5: open the file `mapred-site.xml` C:\Hadoop\hadoop-3.3.3\etc\hadoop

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at
```

```
http://www.apache.org/licenses/LICENSE-2.0
```

```
Unless required by applicable law or agreed to in writing, software  
distributed under the License is distributed on an "AS IS" BASIS,  
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.  
See the License for the specific language governing permissions and  
limitations under the License. See accompanying LICENSE file.
```

```
-->
```

```
<!-- Put site-specific property overrides in this file. -->
```

```
<configuration>
```

```
<property>
```

```
<name>mapreduce.framework.name</name>
```

```
<value>yarn</value>
```

```
</property>
```

```
<property>
```

```
<name>mapreduce.cluster.local.dir</name>
```

```
<value>${hadoop.tmp.dir}/mapred/local</value>
```

```
</property>
```

```
</configuration>
```



```
<> hdfs-site.xml <> core-site.xml <> mapred-site.xml X
<> mapred-site.xml
1  <?xml version="1.0"?>
2  <?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
3  <!--
4      Licensed under the Apache License, Version 2.0 (the "License");
5      you may not use this file except in compliance with the License.
6      You may obtain a copy of the License at
7
8      http://www.apache.org/licenses/LICENSE-2.0
9
10     Unless required by applicable law or agreed to in writing, software
11     distributed under the License is distributed on an "AS IS" BASIS,
12     WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
13     See the License for the specific language governing permissions and
14     limitations under the License. See accompanying LICENSE file.
15 -->
16
17 <!-- Put site-specific property overrides in this file. -->
18 <configuration>
19   <property>
20     <name>mapreduce.framework.name</name>
21     <value>yarn</value>
22   </property>
23   <property>
24     <name>mapreduce.cluster.local.dir</name>
25     <value>${hadoop.tmp.dir}/mapred/local</value>
26   </property>
27
28 </configuration>
29
```

Step 6: open the file **yarn-site -site.xml** C:\Hadoop\hadoop-3.3.3\etc\hadoop

```
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

  http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>

<!-- Site specific YARN configuration properties -->
```

```
<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>
```

```
</> hdfs-site.xml    </> core-site.xml    </> mapred-site.xml    </> yarn-site.xml X
</> yarn-site.xml
1  <?xml version="1.0"?>
2  <!--
3      Licensed under the Apache License, Version 2.0 (the "License");
4      you may not use this file except in compliance with the License.
5      You may obtain a copy of the License at
6
7          http://www.apache.org/licenses/LICENSE-2.0
8
9      Unless required by applicable law or agreed to in writing, software
10     distributed under the License is distributed on an "AS IS" BASIS,
11     WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
12     See the License for the specific language governing permissions and
13     limitations under the License. See accompanying LICENSE file.
14 -->
15 <configuration>
16
17 <!-- Site specific YARN configuration properties -->
18 <configuration>
19 <property>
20 <name>yarn.nodemanager.aux-services</name>
21 <value>mapreduce_shuffle</value>
22 </property>
23 <property>
24 <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
25 <value>org.apache.hadoop.mapred.ShuffleHandler</value>
26 </property>
27
28 </configuration>
29
```

Roll Number: 22-15405

MSC COMPUTER SCIENCE

Subject: Big Data

Step 7: Go to GitHub <https://github.com/Selfgrowth/Apache-hadoop-3.1.1-winutils> and

download the bin folder and replace all the files with the C:\Hadoop\hadoop-3.3.3\bin

Step 8: Open the command prompt and change the location to C:\Hadoop\hadoop-3.3.3\bin

Step 9: hdfs namenode -format

```
C:\Windows\System32\cmd.exe - hdfs namenode -format
Microsoft Windows [Version 10.0.19H4.1826]
(c) Microsoft Corporation. All rights reserved.

C:\Hadoop>hadoop-3.3.3\bin\hdfs namenode -format
2022-07-27 14:56:55,324 INFO namenode.NameNode: STARTUP_MSG:
*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = DESKTOP-D4U1K1L/172.28.48.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 3.3.3
STARTUP_MSG: classpath = C:\Hadoop\hadoop-3.3.3\etc\hadoop:C:\Hadoop\hadoop-3.3.3\share/hadoop/common.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop/common/lib/accessors-smart-2.4.7.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop/common/lib/animal-sniffer-annotations-1.17.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop/common/lib/asn-5.0.4.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop/common/lib/audience-annotations-0.5.0.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop/common/lib/avro-1.7.7.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop/common/lib/checker-qual-2.5.2.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop/common/lib/commons-beanutils-1.9.4.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop/common/lib/commons-cli-1.2.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop/common/lib/commons-codec-1.15.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop/common/lib/commons-collections-3.2.2.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop/common/lib/commons-compress-1.21.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop/common/lib/commons-configuration2-2.11.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common-3.3.3\share/hadoop-3.3.3\share/hadoop-common/lib/commons-io-2.8.0.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/commons-lang-3.12.0.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/commons-logging-1.4.3.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/commons-math3-3.1.1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/commons-net-3.6.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/commons-text-1.4.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/curator-client-4.2.0.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/curator-framework-4.2.0.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/curator-recipes-4.2.0.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/dnsjava-2.1.7.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/failureaccess-1.0.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/gson-2.8.9.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/guava-27.0-jre.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/hadoop-annotations-3.3.3.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/hadoop-auth-3.3.3.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/hadoop-shaded-guava-1.1.1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/hadoop-shaded-protoBuf-3.7.1-1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/httpclient-4.5.13.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/httpcore-4.4.11.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jackson-2.12.0.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jackson-core-2.12.2.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jackson-databind-2.12.2.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jackson-jaxrs-1.9.13.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jackson-mapper-asl-1.9.13.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jakarta.activation-api-1.2.1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/java-xml-jaxb-impl-2.2.3-1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jcip-annotations-1.0-1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jersey-core-1.19.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jersey-json-1.19.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jersey-server-1.19.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jetty-http-9.4.43.v20210629.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jetty-io-9.4.43.v20210629.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jetty-servlet-9.4.43.v20210629.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jetty-util-9.4.43.v20210629.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jetty-webapp-9.4.43.v20210629.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jetty-xml-9.4.43.v20210629.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jsch-0.1.55.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/json-smart-2.4.7.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jsp-api-2.1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jsr305-3.0.2.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jsr311-api-1.1.1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/jul-to-slf4j-1.7.36.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/kerb-admin-1.0.1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/kerb-client-1.0.1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/kerb-common-1.0.1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/kerb-crypto-1.0.1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/kerb-idnt-1.0.1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/kerb-rt-1.0.1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/kerby-asn1-1.0.1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/kerby-config-1.0.1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/kerby-dkix-1.0.1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/kerby-util-1.0.1.jar;C:\Hadoop\hadoop-3.3.3\share/hadoop-common/lib/kerby-x509-1.0.1.jar
```

```
C:\Windows\System32\cmd.exe
2022-07-27 14:56:57,410 INFO util.GSet: Computing capacity for map InodeMap
2022-07-27 14:56:57,410 INFO util.GSet: VM type = 64-bit
2022-07-27 14:56:57,413 INFO util.GSet: 1.0% max memory 889 MB = 8.9 MB
2022-07-27 14:56:57,415 INFO util.GSet: capacity = 2^20 = 1048576 entries
2022-07-27 14:56:57,417 INFO namenode.FSDirectory: ACLs enabled? true
2022-07-27 14:56:57,417 INFO namenode.FSDirectory: POSIX ACL inheritance enabled? true
2022-07-27 14:56:57,419 INFO namenode.FSDirectory: XAttrs enabled? true
2022-07-27 14:56:57,420 INFO namenode.NameNode: Caching file names occurring more than 10 times
2022-07-27 14:56:57,433 INFO snapshot.SnapshotManager: Loaded config captureOpenfiles: false, skipCaptureAccessTimeOnlyChange: false, snapshotDiffAllowSnapRootDescendant: true, maxSnapshotLimit: 65536
2022-07-27 14:56:57,438 INFO snapshot.SnapshotManager: Skiplist is disabled
2022-07-27 14:56:57,448 INFO util.GSet: Computing capacity for map cachedBlocks
2022-07-27 14:56:57,448 INFO util.GSet: VM type = 64-bit
2022-07-27 14:56:57,452 INFO util.GSet: 0.25% max memory 889 MB = 2.2 MB
2022-07-27 14:56:57,452 INFO util.GSet: capacity = 2^18 = 262144 entries
2022-07-27 14:56:57,475 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2022-07-27 14:56:57,476 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2022-07-27 14:56:57,476 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2022-07-27 14:56:57,490 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2022-07-27 14:56:57,492 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
2022-07-27 14:56:57,497 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2022-07-27 14:56:57,497 INFO util.GSet: VM type = 64-bit
2022-07-27 14:56:57,499 INFO util.GSet: 0.029999999329447746% max memory 889 MB = 273.1 KB
2022-07-27 14:56:57,501 INFO util.GSet: capacity = 2^15 = 32768 entries
Re-format filesystem in Storage Directory root= C:\Hadoop\hadoop-3.3.3\DATA\NameNode; location= null ? (Y or N) Y
2022-07-27 14:58:09,984 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1977693252-172.28.48.1-1658914889970
2022-07-27 14:58:09,985 INFO common.Storage: Will remove files: [C:\Hadoop\hadoop-3.3.3\DATA\NameNode\current\fsimage_000000000000000000, C:\Hadoop\hadoop-3.3.3\DATA\NameNode\current\fsimage_000000000000000000.mds, C:\Hadoop\hadoop-3.3.3\DATA\NameNode\current\seen_txid, C:\Hadoop\hadoop-3.3.3\DATA\NameNode\current\VERSION]
2022-07-27 14:58:09,986 INFO common.Storage: Storage directory C:\Hadoop\hadoop-3.3.3\DATA\NameNode has been successfully formatted.
2022-07-27 14:58:10,081 INFO namenode.FSImageFormatProtobuf: Saving image file C:\Hadoop\hadoop-3.3.3\DATA\NameNode\current\fsimage_000000000000000000 using no compression
2022-07-27 14:58:10,296 INFO namenode.FSImageFormatProtobuf: Image file C:\Hadoop\hadoop-3.3.3\DATA\NameNode\current\fsimage.ckpt_000000000000000000 of size 399 bytes saved in 0 seconds .
2022-07-27 14:58:10,310 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2022-07-27 14:58:10,341 INFO namenode.FSNamesystem: Stopping services started for active state
2022-07-27 14:58:10,341 INFO namenode.FSNamesystem: Stopping services started for standby state
2022-07-27 14:58:10,346 INFO namenode.FSImage: FSImageSaver: clean checkpoint: txid=0 when meet shutdown.
2022-07-27 14:58:10,346 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at DESKTOP-D4U1K1L/172.28.48.1
*****/
C:\Hadoop\hadoop-3.3.3\bin>
```

Step 10: Go to sbin folder. Type cmd in the address bar and 2 different open command prompt

First cmd: **.\start-dfs.cmd**

```
Select C:\Windows\System32\cmd.exe
Microsoft Windows [Version 10.0.19044.1826]
(c) Microsoft Corporation. All rights reserved.

C:\Hadoop\hadoop-3.3.3\sbin>.start-dfs.cmd
C:\Hadoop\hadoop-3.3.3\sbin>_
```

Second cmd: **.\start-yarn.cmd**

```
C:\Windows\System32\cmd.exe
Microsoft Windows [Version 10.0.19044.1826]
(c) Microsoft Corporation. All rights reserved.

C:\Hadoop\hadoop-3.3.3\sbin>.start-yarn.cmd
starting yarn daemons
C:\Hadoop\hadoop-3.3.3\sbin>_
```

Step 11: now open chrome or any other browser and type <http://localhost:9870/>

The screenshot shows a web browser window with the URL <http://localhost:9870/dfshealth.html#tab-overview>. The page title is "NameNode information". The main content area has a green header with tabs: "Hadoop", "Overview", "Datanodes", "Datanode Volume Failures", "Snapshot", "Startup Progress", and "Utilities". The "Overview" tab is selected, showing "Overview 'localhost:9820' (✓active)". Below this is a table with the following information:

Started:	Wed Jul 27 15:00:02 +0530 2022
Version:	3.3.3, rd37586cbda38c338d9fe481adda5a05fb516771
Compiled:	Mon May 09 22:06:00 +0530 2022 by steevel from branch-3.3.3
Cluster ID:	CID-421309ee-a178-42fc-a6a4-ec5c50084dce
Block Pool ID:	BP-1977693252-172.28.48.1-1658914089970

Below the table is a "Summary" section with the following text:

Security is off.
Safemode is off.
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).
Heap Memory used 70.68 MB of 191.5 MB Heap Memory. Max Heap Memory is 889 MB.

Practical 3

Aim: Write an Hadoop MapReduce Program in Python

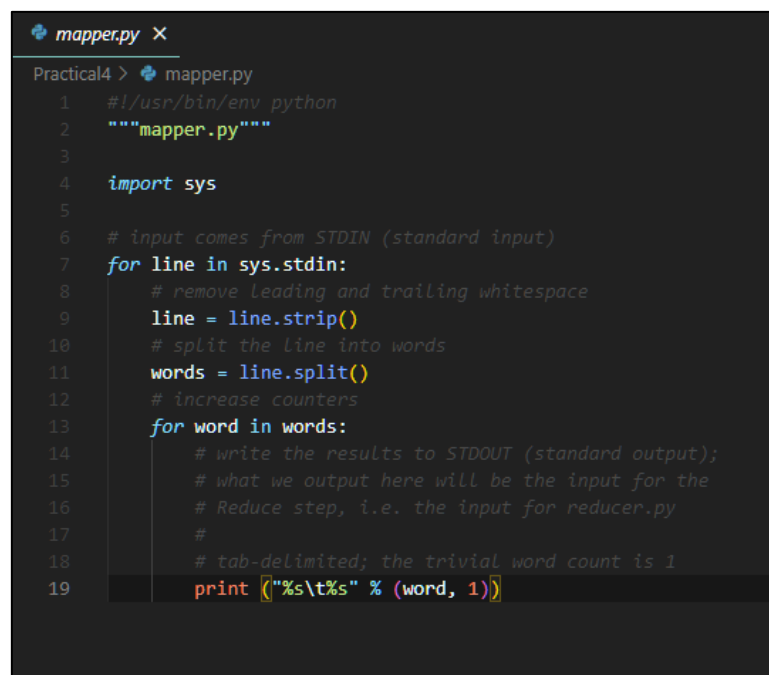


Create the mapper.py

```
#!/usr/bin/env python
"""mapper.py"""

import sys

# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # split the line into words
    words = line.split()
    # increase counters
    for word in words:
        # write the results to STDOUT (standard output);
        # what we output here will be the input for the
        # Reduce step, i.e. the input for reducer.py
        #
        # tab-delimited; the trivial word count is 1
        print ("%s\t%s" % (word, 1))
```



```
mapper.py X
Practical4 > mapper.py
1  #!/usr/bin/env python
2  """mapper.py"""
3
4  import sys
5
6  # input comes from STDIN (standard input)
7  for line in sys.stdin:
8      # remove leading and trailing whitespace
9      line = line.strip()
10     # split the line into words
11     words = line.split()
12     # increase counters
13     for word in words:
14         # write the results to STDOUT (standard output);
15         # what we output here will be the input for the
16         # Reduce step, i.e. the input for reducer.py
17         #
18         # tab-delimited; the trivial word count is 1
19         print ("%s\t%s" % (word, 1))
```

Create the reducer.py

```
#!/usr/bin/env python
"""reducer.py"""

from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None

# input comes from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()

    # parse the input we got from mapper.py
    word, count = line.split('\t', 1)

    # convert count (currently a string) to int
    try:
        count = int(count)
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue

    # this IF-switch only works because Hadoop sorts map output
    # by key (here: word) before it is passed to the reducer
    if current_word == word:
        current_count += count
    else:
        if current_word:
            # write result to STDOUT
            print ("%s\t%s" % (current_word, current_count))
        current_count = count
        current_word = word

# do not forget to output the last word if needed!
if current_word == word:
    print ("%s\t%s" % (current_word, current_count))
```

```
reducer.py X
Practical4 > reducer.py
1  #!/usr/bin/env python
2  """reducer.py"""
3
4  from operator import itemgetter
5  import sys
6
7  current_word = None
8  current_count = 0
9  word = None
10
11 # input comes from STDIN
12 for line in sys.stdin:
13     # remove leading and trailing whitespace
14     line = line.strip()
15
16     # parse the input we got from mapper.py
17     word, count = line.split('\t', 1)
18
19     # convert count (currently a string) to int
20     try:
21         count = int(count)
22     except ValueError:
23         # count was not a number, so silently
24         # ignore/discard this line
25         continue
26
27     # this IF-switch only works because Hadoop sorts map output
28     # by key (here: word) before it is passed to the reducer
29     if current_word == word:
30         current_count += count
31     else:
32         if current_word:
33             # write result to STDOUT
34             print ("%s\t%s" % (current_word, current_count))
35             current_count = count
36             current_word = word
37
38 # do not forget to output the last word if needed!
39 if current_word == word:
40     print ("%s\t%s" % (current_word, current_count))
```

Running mapper and reducer without Hadoop HDFS

Step 1: Open **Command Prompt** where the mapper.py and reducer.py is located

```
PROBLEMS  OUTPUT  TERMINAL  JUPYTER  COMMENTS  DEBUG CONSOLE
Microsoft Windows [Version 10.0.19044.1826]
(c) Microsoft Corporation. All rights reserved.

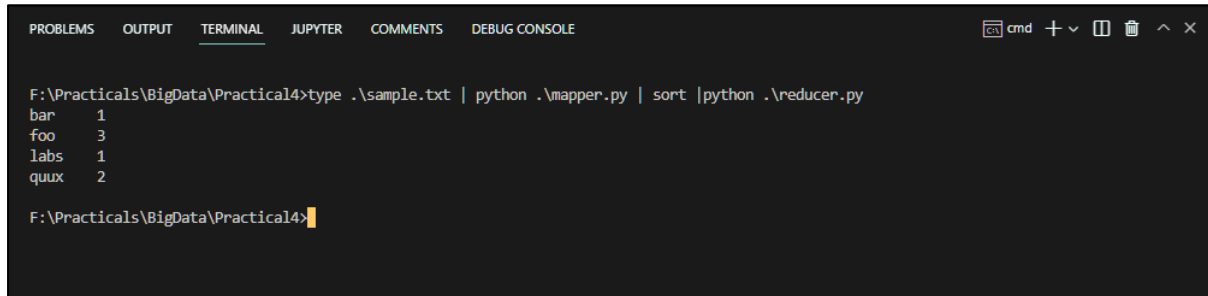
F:\Practicals\BigData>cd Practical4
F:\Practicals\BigData\Practical4>
```

Step 2: To Execute the program create one file in the same location with name sample.txt

```
sample.txt X
Practical4 > sample.txt
1  foo foo quux labs foo bar quux
```

Step 3: Now run the following command to get the output

```
type .\sample.txt | python .\mapper.py | sort | python .\reducer.py
```



```
PROBLEMS OUTPUT TERMINAL JUPYTER COMMENTS DEBUG CONSOLE cmd + - [ ] [ ] ^ X

F:\Practicals\BigData\Practical4>type .\sample.txt | python .\mapper.py | sort | python .\reducer.py
bar      1
foo      3
labs     1
quux     2

F:\Practicals\BigData\Practical4>
```

Running the Python Code on Hadoop

Step 1: Download example input data

We will use three eBooks from Project Gutenberg for this example:

- [The Outline of Science, Vol. 1 \(of 4\) by J. Arthur Thomson](#)
- [The Notebooks of Leonardo Da Vinci](#)
- [Ulysses by James Joyce](#)

Download each eBook as text files in `Plain Text UTF-8` encoding and store the files in a local temporary directory of choice.

Copy local example data to HDFS

Before we run the actual MapReduce job, we must first copy the files from our local file system to Hadoop's HDFS.

Step 1: Open Command Prompt in Administration Mode and change the present working directory to the `C:\Hadoop\hadoop-3.3.3\sbin`


```
Administrator: Command Prompt
Microsoft Windows [Version 10.0.19044.1826]
(c) Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>cd ..

C:\Windows> cd ..

C:\> cd Hadoop

C:\Hadoop>cd hadoop-3.3.3

C:\Hadoop\hadoop-3.3.3> cd sbin

C:\Hadoop\hadoop-3.3.3\sbin>
```

Step 2: Now run the command `.\start-all.cmd`

```
Administrator: Command Prompt
Microsoft Windows [Version 10.0.19044.1826]
(c) Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>cd ..

C:\Windows> cd ..

C:\> cd Hadoop

C:\Hadoop>cd hadoop-3.3.3

C:\Hadoop\hadoop-3.3.3> cd sbin

C:\Hadoop\hadoop-3.3.3\sbin>.\start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\Hadoop\hadoop-3.3.3\sbin>
```

Step3: Now change the present working directory to `C:\Hadoop\hadoop-3.3.3\bin` and run the command

`hadoop dfs -copyFromLocal 'path of the downloaded sample file' 'path to store on the hdfs'`

```
hadoop dfs -copyFromLocal
"F:\Practicals\BigData\Practical4\Data" hdfs://localhost:9000/Harsh
```

```
Administrator: Command Prompt
C:\Hadoop\hadoop-3.3.3\sbin>cd ..
C:\Hadoop\hadoop-3.3.3> cd bin
C:\Hadoop\hadoop-3.3.3\bin> hadoop dfs -copyFromLocal "F:\Practicals\BigData\Practical4\Data" hdfs://localhost:9000/Harsh
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
C:\Hadoop\hadoop-3.3.3\bin>_
```

```
hadoop dfs -ls /Harsh
```

```
Administrator: Command Prompt
C:\Hadoop\hadoop-3.3.3\bin>hadoop dfs -ls /Harsh
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
Found 4 items:
-rw-r--r-- 3 User supergroup 1586331 2022-07-29 09:32 /Harsh/4300-0.txt
-rw-r--r-- 3 User supergroup 1428909 2022-07-29 09:32 /Harsh/5000-8.txt
-rw-r--r-- 3 User supergroup 674565 2022-07-29 09:32 /Harsh/pg20417.txt
-rw-r--r-- 3 User supergroup 30 2022-07-29 09:32 /Harsh/sample.txt
C:\Hadoop\hadoop-3.3.3\bin>
```

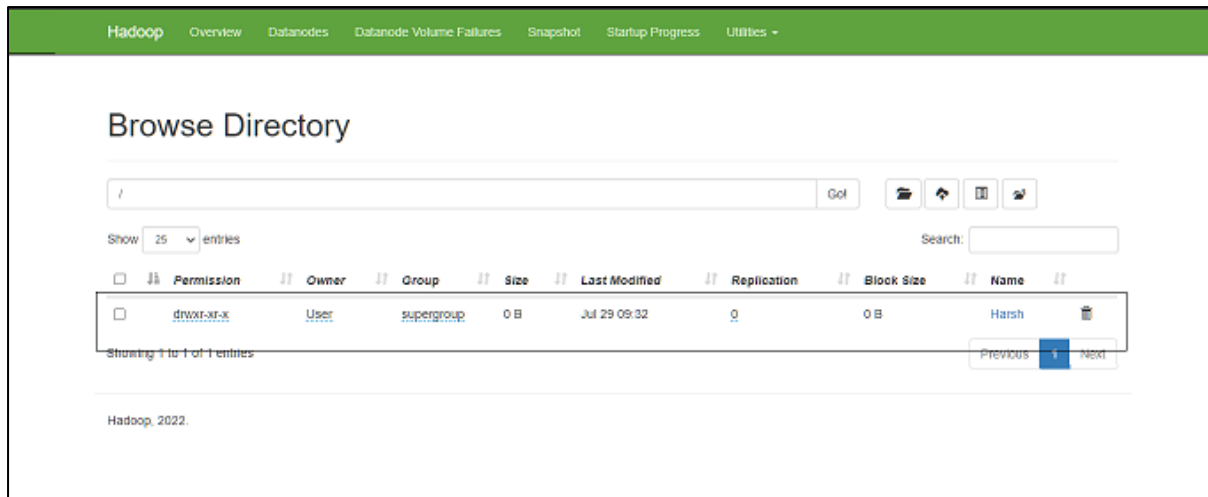
Step 4: To check the files are uploaded to the Hadoop HDFS the visit <http://localhost:9870/dfshealth.html#tab-overview> → go to utilities in the navigation bar and click on the Browse the file system

The screenshot shows the Hadoop Overview page for 'localhost:9000' (active). The 'Utilities' dropdown menu is open, highlighting 'Browse the file system'. Below the menu, a table displays cluster metadata.

Property	Value
Started:	Fri Jul 29 09:27:02 +0530 2022
Version:	3.3.3, rd37586cbda38c338d9fe481adda5a05fb0167f1
Compiled:	Mon May 09 22:06:00 +0530 2022 by stevel from branch-3.3.3
Cluster ID:	CID-421309ee-a178-42fc-a6a4-ec5c50084dce
Block Pool ID:	BP-1977693252-172.28.48.1-1658914089970

Summary

Security is off.
Safemode is off.
6 files and directories, 4 blocks (4 replicated blocks, 0 erasure coded block groups) = 10 total filesystem object(s).



Step 5: Run the MapReduce job

```
hadoop jar C:\Hadoop\hadoop-3.3.3\share\hadoop\tools\lib\hadoop-streaming-3.3.3.jar -file F:\Practicals\BigData\Practical4\mapper.py -mapper "python mapper.py" -file F:\Practicals\BigData\Practical4\reducer.py -reducer "python reducer.py" -input hdfs://localhost:9000/Harsh/sample.txt -output /output
```

```
Administrator: Command Prompt
C:\Hadoop\hadoop-3.3.3>cd ..
C:\Hadoop\hadoop-3.3.3> cd bin
C:\Hadoop\hadoop-3.3.3\bin>hadoop dfs -copyFromLocal "F:\Practicals\BigData\Practical4\Data" hdfs://localhost:9000/Harsh
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
C:\Hadoop\hadoop-3.3.3\bin>hadoop jar C:\Hadoop\hadoop-3.3.3\share\hadoop\tools\lib\hadoop-streaming-3.3.3.jar -file F:\Practicals\BigData\Practical4\mapper.py -mapper "python mapper.py" -file F:\Practicals\BigData\Practical4\reducer.py -reducer "python reducer.py" -input hdfs://localhost:9000/Harsh/sample.txt -output /output
```

```
Administrator: Command Prompt
C:\Hadoop\hadoop-3.3.3\bin>hadoop jar C:\Hadoop\hadoop-3.3.3\share\hadoop\tools\lib\hadoop-streaming-3.3.3.jar -file F:\Practicals\BigData\Practical4\mapper.py -mapper "python mapper.py" -file F:\Practicals\BigData\Practical4\reducer.py -reducer "python reducer.py" -input hdfs://localhost:9000/Harsh/sample.txt -output /output
packageJobJar: [F:\Practicals\BigData\Practical4\mapper.py, F:\Practicals\BigData\Practical4\reducer.py, /C:/Users/Admin/AppData/Local/Temp/hadoop-unjar7499095751082734589/] [] C:\Users\Admin\AppData\Local\Temp\streamjob1790884674952075886.jar tmpDir=null
2022-07-29 09:41:05,696 INFO client.DefaultHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-07-29 09:41:06,574 INFO client.DefaultHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-07-29 09:41:08,481 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-user/staging/User/.staging/job_1659067024247_0001
2022-07-29 09:41:09,402 INFO mapred.FileInputFormat: Total input files to process : 1
2022-07-29 09:41:09,500 INFO mapreduce.JobSubmitter: number of splits:2
2022-07-29 09:41:10,069 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1659067024247_0001
2022-07-29 09:41:10,069 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-07-29 09:41:10,475 INFO conf.Configuration: resource-types.xml not found
2022-07-29 09:41:10,477 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-07-29 09:41:12,197 INFO impl.YarnClientImpl: Submitted application application_1659067024247_0001
2022-07-29 09:41:12,370 INFO mapreduce.Job: The url to track the job: http://DESKTOP-D4U1K1L:8088/proxy/application_1659067024247_0001/
2022-07-29 09:41:12,457 INFO mapreduce.Job: Running job: job_1659067024247_0001
2022-07-29 09:41:50,560 INFO mapreduce.Job: Job job_1659067024247_0001 running in uber mode : false
2022-07-29 09:41:50,575 INFO mapreduce.Job: map 0% reduce 0%
2022-07-29 09:42:20,042 INFO mapreduce.Job: map 100% reduce 0%
2022-07-29 09:42:40,633 INFO mapreduce.Job: map 100% reduce 100%
2022-07-29 09:42:41,665 INFO mapreduce.Job: Job job_1659067024247_0001 completed successfully
2022-07-29 09:42:41,937 INFO mapreduce.Job: Counters: 54
File System Counters
  FILE: Number of bytes read=65
  FILE: Number of bytes written=842595
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=225
  HDFS: Number of bytes written=26
  HDFS: Number of read operations=11
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
  HDFS: Number of bytes read erasure-coded=0
Job Counters
  Launched map tasks=2
  Launched reduce tasks=1
  Rack-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=51238
  Total time spent by all reduces in occupied slots (ms)=16225
  Total time spent by all map tasks (ms)=51238
  Total time spent by all reduce tasks (ms)=16225
  Total vcore-milliseconds taken by all map tasks=51238
```

```
Administrator: Command Prompt
Total vcore-milliseconds taken by all map tasks=51238
Total vcore-milliseconds taken by all reduce tasks=16225
Total megabyte-milliseconds taken by all map tasks=52467712
Total megabyte-milliseconds taken by all reduce tasks=16614400
Map-Reduce Framework
  Map input records=1
  Map output records=7
  Map output bytes=45
  Map output materialized bytes=71
  Input split bytes=180
  Combine input records=0
  Combine output records=0
  Reduce input groups=4
  Reduce shuffle bytes=71
  Reduce input records=7
  Reduce output records=4
  Spilled Records=14
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=368
  CPU time spent (ms)=4246
  Physical memory (bytes) snapshot=752508928
  Virtual memory (bytes) snapshot=1077448704
  Total committed heap usage (bytes)=630718464
  Peak Map Physical memory (bytes)=271998784
  Peak Map Virtual memory (bytes)=379629568
  Peak Reduce Physical memory (bytes)=209387520
  Peak Reduce Virtual memory (bytes)=321359872
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=45
File Output Format Counters
  Bytes Written=26
2022-07-29 09:42:41,940 INFO streaming.StreamJob: Output directory: /output
C:\Hadoop\hadoop-3.3.3\bin>
```

Step 6: Check if the result is successfully stored in HDFS directory `/output`

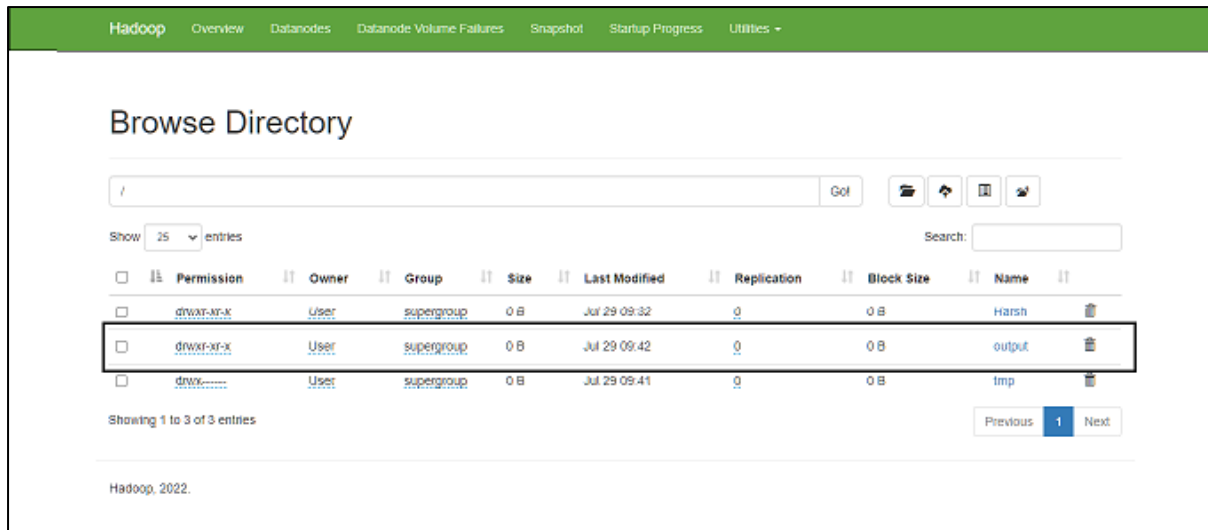
```
hadoop dfs -ls /output
```

```
Administrator: Command Prompt
C:\Hadoop\hadoop-3.3.3\bin>hadoop dfs -ls /output
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
Found 2 items
-rw-r--r--  3 User supergroup      0 2022-07-29 09:42 /output/_SUCCESS
-rw-r--r--  3 User supergroup    26 2022-07-29 09:42 /output/part-00000
C:\Hadoop\hadoop-3.3.3\bin>
```

Step 7: To check the output is generated to the Hadoop HDFS the visit <http://localhost:9870/dfshealth.html#tab-overview> → go to utilities in the navigation bar and click on the Browse the file system

The screenshot shows the Hadoop Overview page for 'localhost:9000' (active). The 'Utilities' menu is open, showing options like 'Browse the file system', 'Logs', 'Log Level', 'Metrics', 'Configuration', 'Process Thread Dump', and 'Network Topology'. The 'Browse the file system' option is highlighted. Below the menu, the 'Overview' section displays a table with details about the Hadoop cluster, including 'Started', 'Version', 'Compiled', 'Cluster ID', and 'Block Pool ID'. The 'Summary' section below shows that security and safemode are off, and there are 6 files and directories, 4 blocks (4 replicated blocks, 0 erasure coded block groups) = 10 total filesystem object(s).

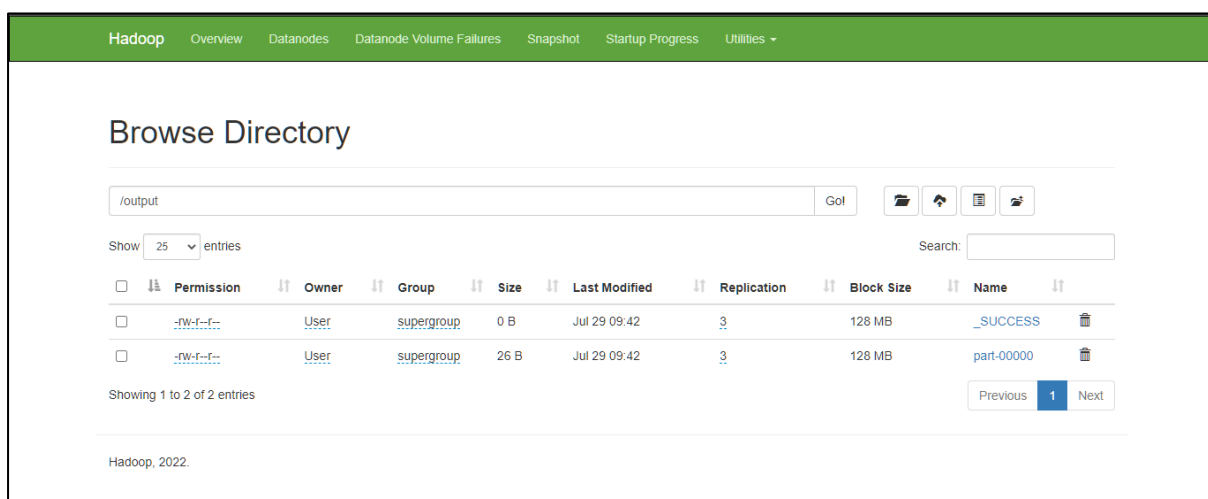
Started:	Fri Jul 29 09:27:02 +0530 2022
Version:	3.3.3, rd37585cbda38c338d9fe481addda5a05fb61671
Compiled:	Mon May 09 22:06:00 +0530 2022 by steele from branch-3.3.3
Cluster ID:	CID-421309ee-a178-42fc-a694-ec5c50084dce
Block Pool ID:	BP-1977693252-172.28.48.1-1658914089970



Hadoop Browse Directory interface showing the root directory (/). The table lists three entries:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	User	supergroup	0 B	Jul 29 09:42	0	0 B	Harsh
drwxr-xr-x	User	supergroup	0 B	Jul 29 09:42	0	0 B	output
drwxr-xr-x	User	supergroup	0 B	Jul 29 09:41	0	0 B	tmp

Showing 1 to 3 of 3 entries



Hadoop Browse Directory interface showing the contents of the /output directory. The table lists two entries:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	User	supergroup	0 B	Jul 29 09:42	3	128 MB	_SUCCESS
-rw-r--r--	User	supergroup	26 B	Jul 29 09:42	3	128 MB	part-00000

Showing 1 to 2 of 2 entries

Step 8: You can then inspect the contents of the file with the `fs -cat` command:

```
hadoop fs -cat /output/part-00000
```



```
Administrator: Command Prompt
C:\Hadoop\hadoop-3.3.3\bin> hadoop fs -cat /output/part-00000
bar      1
foo      3
labs     1
quux     2
C:\Hadoop\hadoop-3.3.3\bin>
```