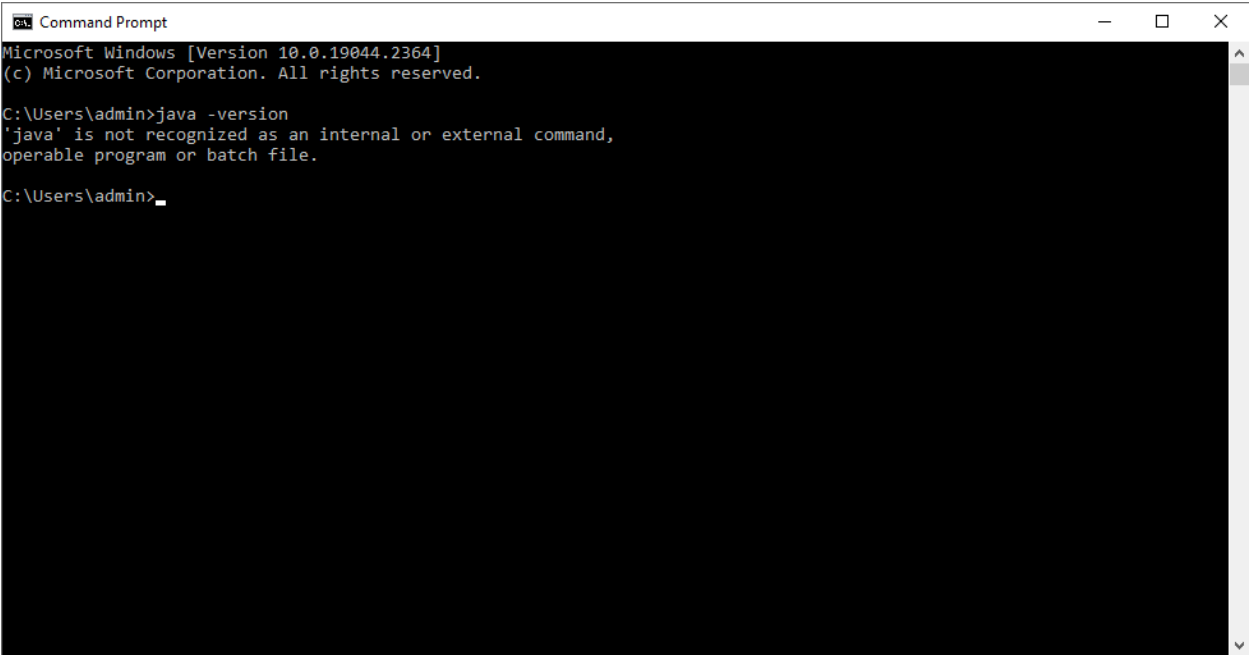M2 – Assignment 1

HADOOP INSTALLATION

Hadoop needs Java to run. Using your command prompt (cmd), check if Java is already installed in your system. Type this command on your cmd: "java -version" to check the installed java version.


The message shows that you don't have a java installed in your device. Therefore, you need to go to this link: https://java.com/en/download/ and download Java. Scroll down and you will see Java 8, choose the appropriate OS and product to your device.
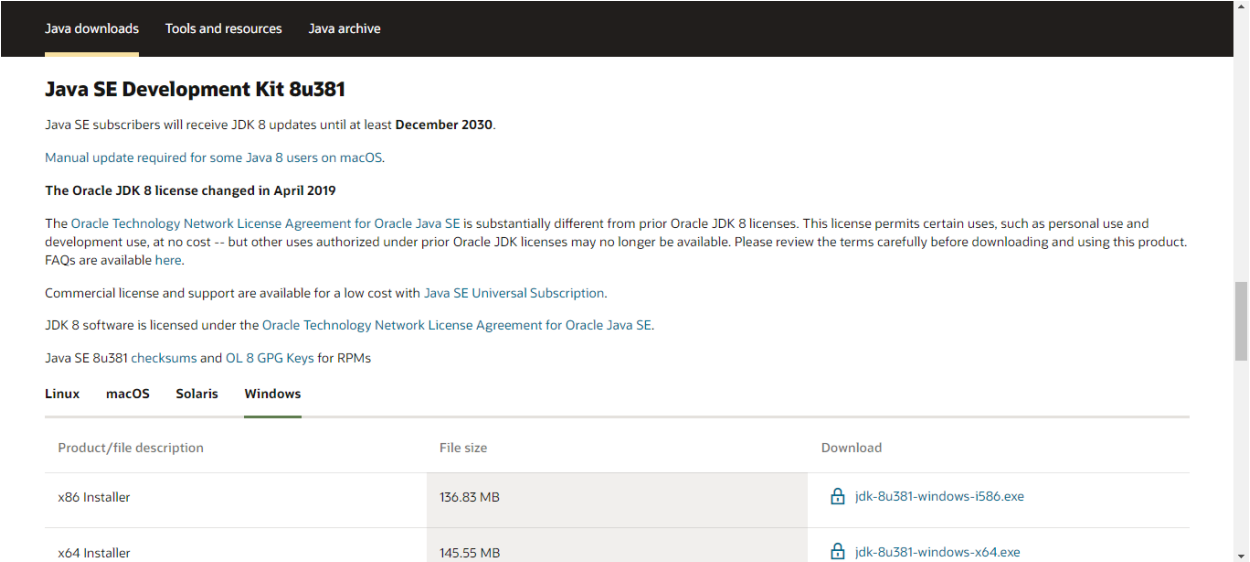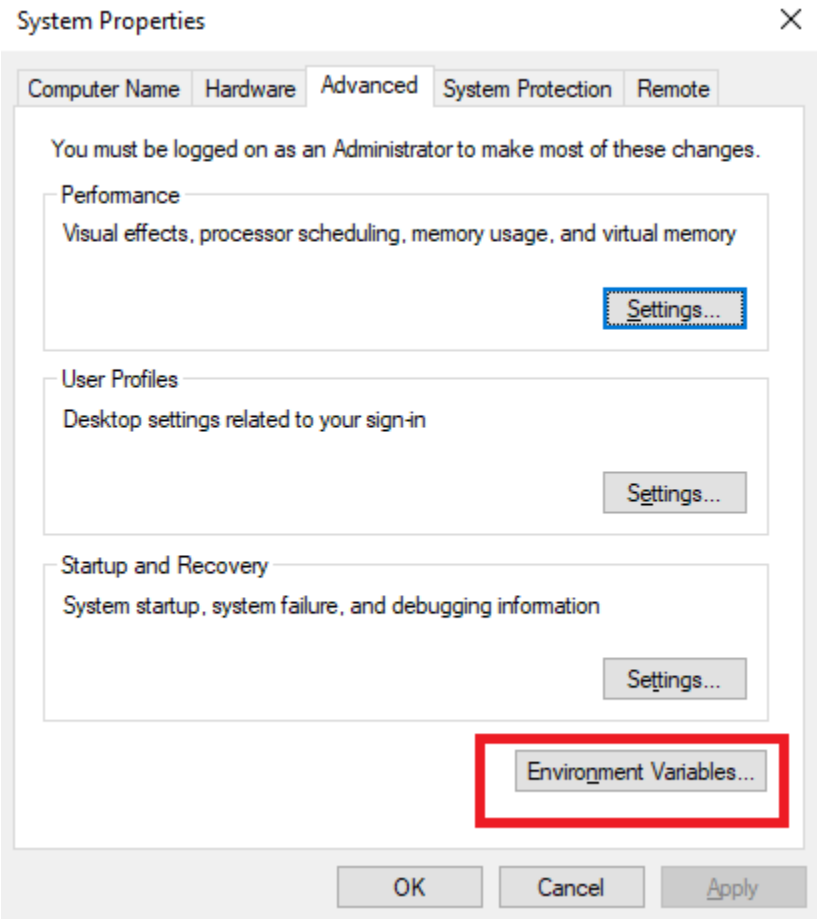

After downloading the file, install the Java.
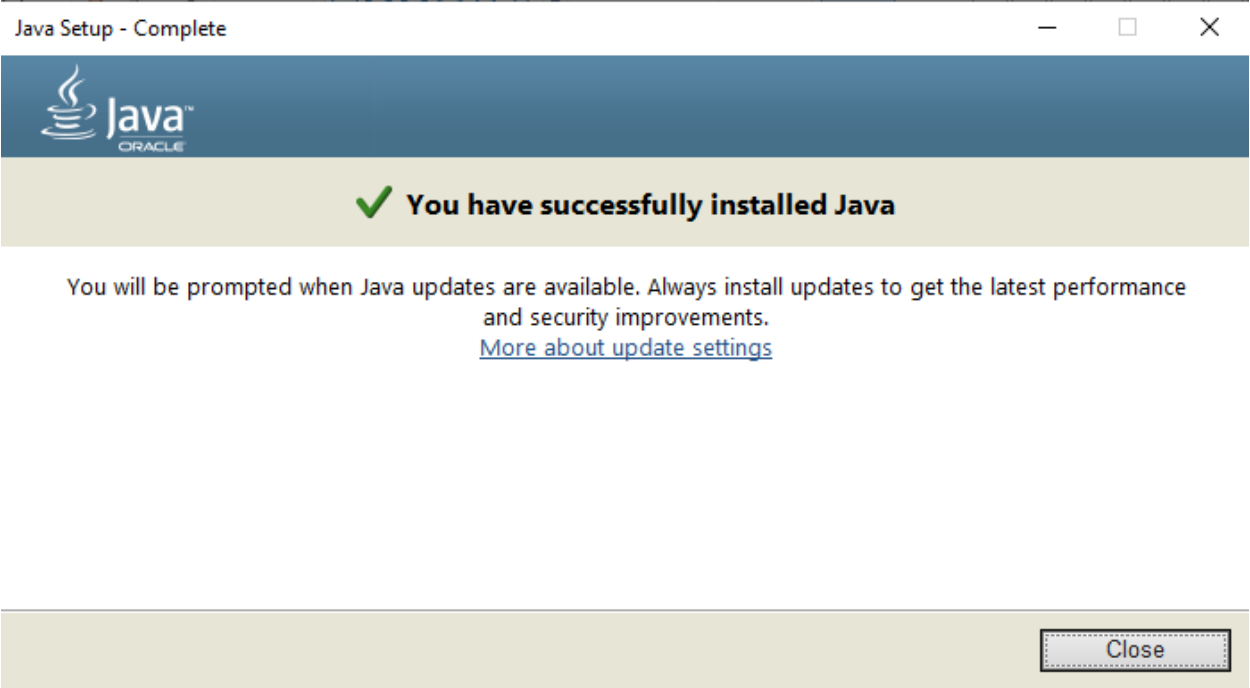
Now, on your windows search bar, type "environment variables", this should appear as shown below and press it.

Edit the system **environment variables**
Control panel
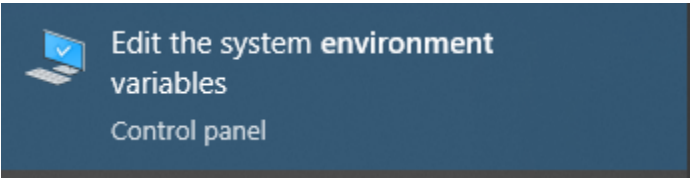
A tab named System Properties will show after clicking it. Click the Environment Variables located on the lower right of the tab.

On Evironmental Variables tab, under the User variables., press New.
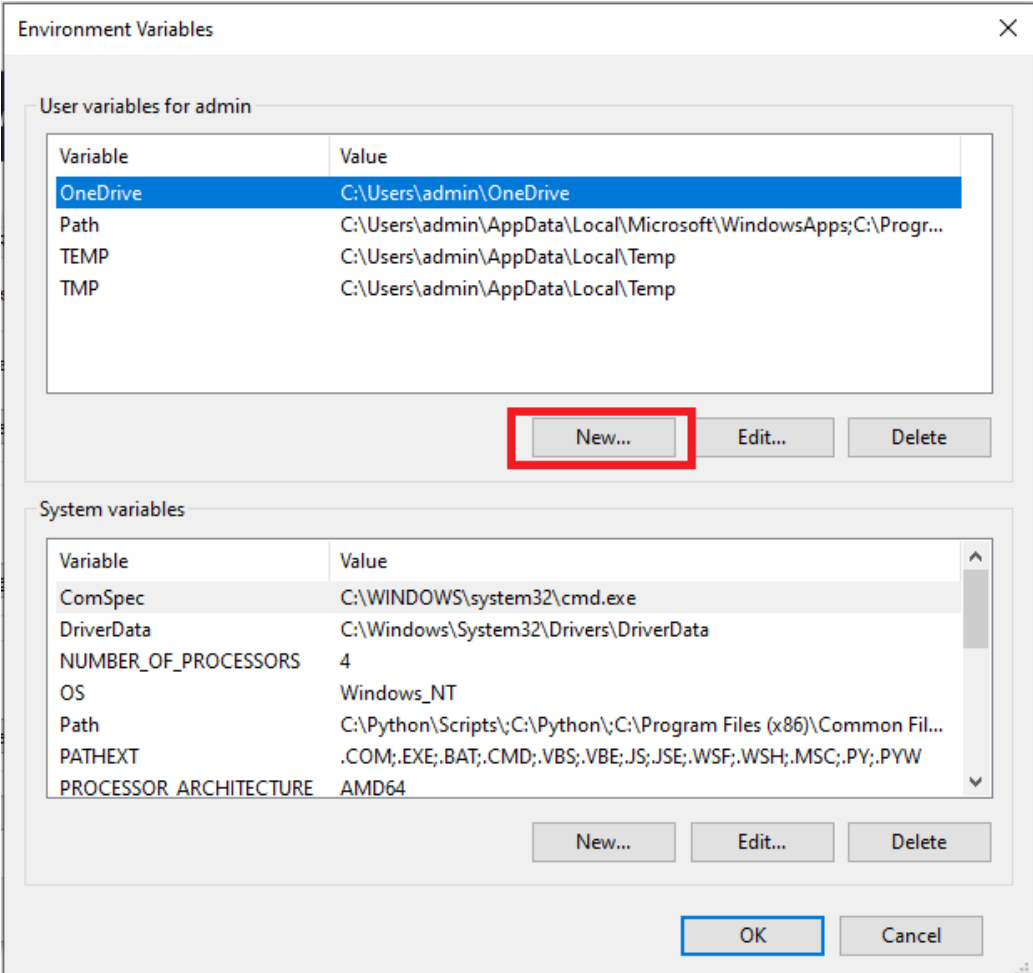
Set "JAVA_HOME" on the variable name. Then, put the file location of the bin folder on your java folder in the variable value.

Now, click the Path on System variables and press Edit.

Click New, and add the directory of the bin folder on your java folder. Then, press OK.

Open cmd again and check your Java version, use the command we provided earlier to check java version..

The Java version now appears. It shows that you have successfully installed Java.

Open this link: https://hadoop.apache.org/releases.html. You should be able to see this site shown below. Click binary.

After clicking the link, this site shown below should appear. Click the first link provided and the download will now start.
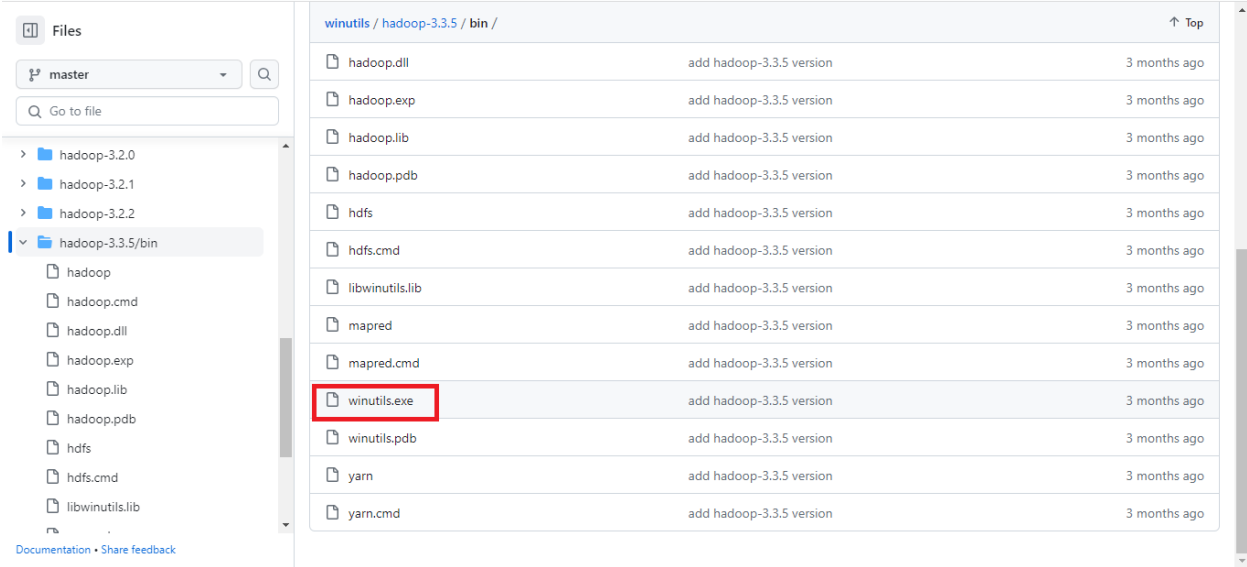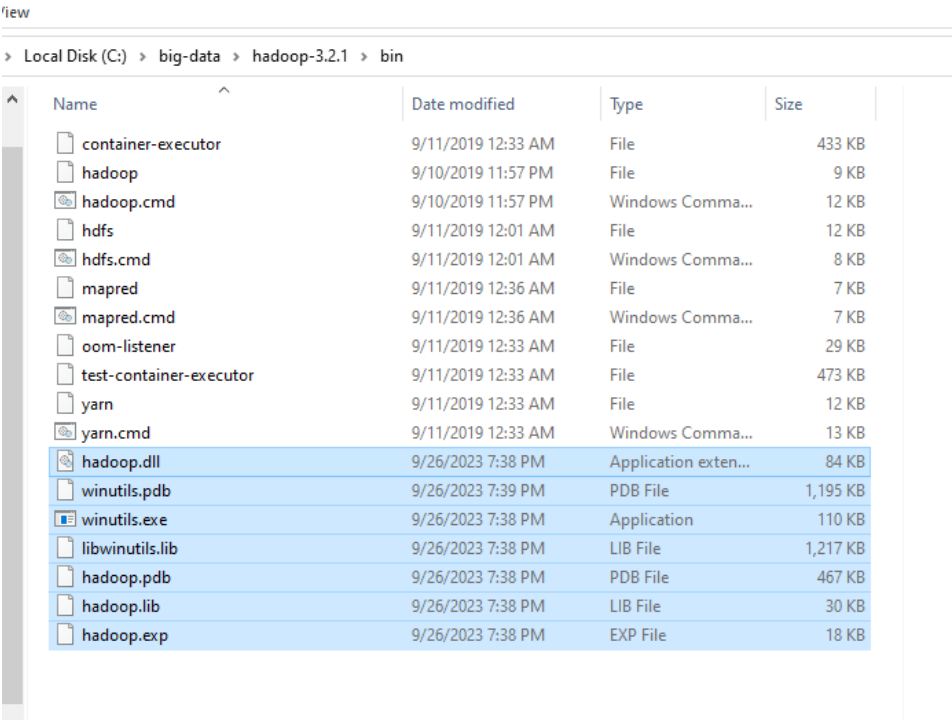

Create a folder you want on your C:\ and extract the downloaded file on the folder.
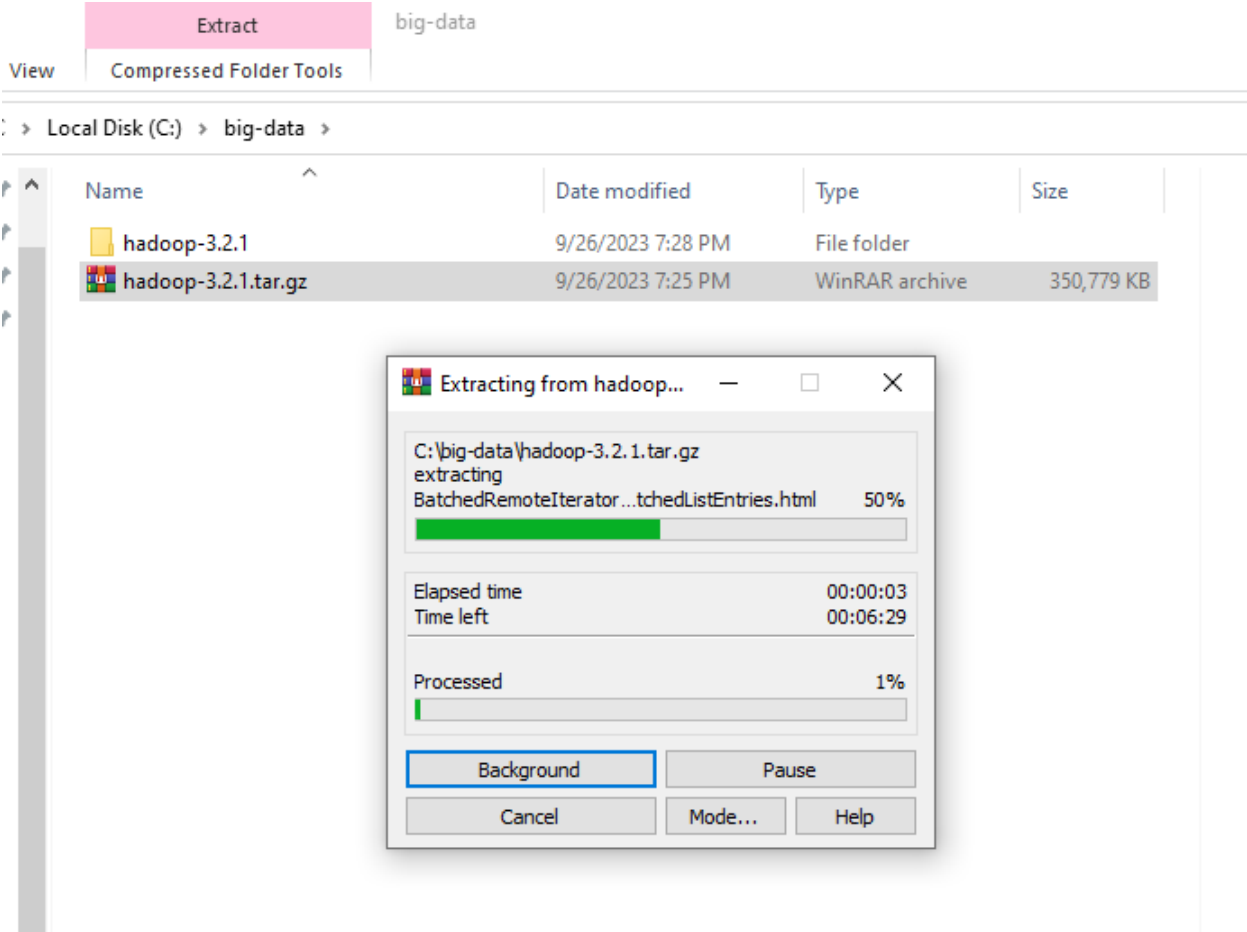
After that, open this link: https://github.com/cdarlint/winutils. Find the version you downloaded and inside the bin folder, download winutils.exe.

On your C:\, create "hadoop" folder and inside it, create a "bin" folder. Put the "winutils.exe" file you downloaded earlier inside the "bin" folder.

Open Environment variables again and create new user variable. Set "HADOOP_HOME" on the variable name. Then, put the bin location on your Hadoop folder in the variable value.

Edit User Variable ✕

Variable name: HADOOP_HOME

Variable value: C:\big-data\hadoop-3.2.1\bin

Browse Directory...   Browse File...   OK   Cancel

Click Path on the system variable and press edit. Click new and put the same directory or the bin folder on your Hadoop folder.

Edit environment variable ✕

C:\Program Files\Common Files\Oracle\Java\javapath
C:\Python27
C:\Python\
C:\Program Files (x86)\Common Files\Oracle\Java\javapath
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
C:\Program Files\Intel\WiFi\bin\
C:\Program Files\Common Files\Intel\WirelessCommon\
C:\Cassandra\apache-cassandra-3.11.16\bin
C:\Python\Scripts
C:\Python27\Tools\Scripts
C:\Program Files\Java\jdk-1.8\bin
C:\big-data\hadoop-3.2.1\bin

New
Edit
Browse...
Delete

Move Up
Move Down

Edit text...

OK   Cancel

After this, check if Hadoop is working. To know that, run your cmd and type hadoop.

```
Command Prompt                                                    —  □  ✕

Microsoft Windows [Version 10.0.19045.3448]
(c) Microsoft Corporation. All rights reserved.

C:\Users\admin>hadoop
Usage: hadoop [--config confdir] [--loglevel loglevel] COMMAND
where COMMAND is one of:
  fs                   run a generic filesystem user client
  version              print the version
  jar <jar>            run a jar file
                       note: please use "yarn jar" to launch
                             YARN applications, not this command.
  checknative [-a|-h]  check native hadoop and compression libraries availability
  conftest             validate configuration XML files
  distch path:owner:group:permisson
                       distributed metadata changer
  distcp <srcurl> <desturl> copy file or directories recursively
  archive -archiveName NAME -p <parent path> <src>* <dest> create a hadoop archive
  classpath            prints the class path needed to get the
                       Hadoop jar and the required libraries
  credential           interact with credential providers
  jnipath              prints the java.library.path
  kerbname             show auth_to_local principal conversion
  kdiag                diagnose kerberos problems
  key                  manage keys via the KeyProvider
  trace                view and modify Hadoop tracing settings
  daemonlog            get/set the log level for each daemon
 or
  CLASSNAME            run the class named CLASSNAME

Most commands print help when invoked w/o parameters.

C:\Users\admin>
```
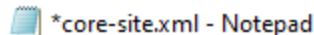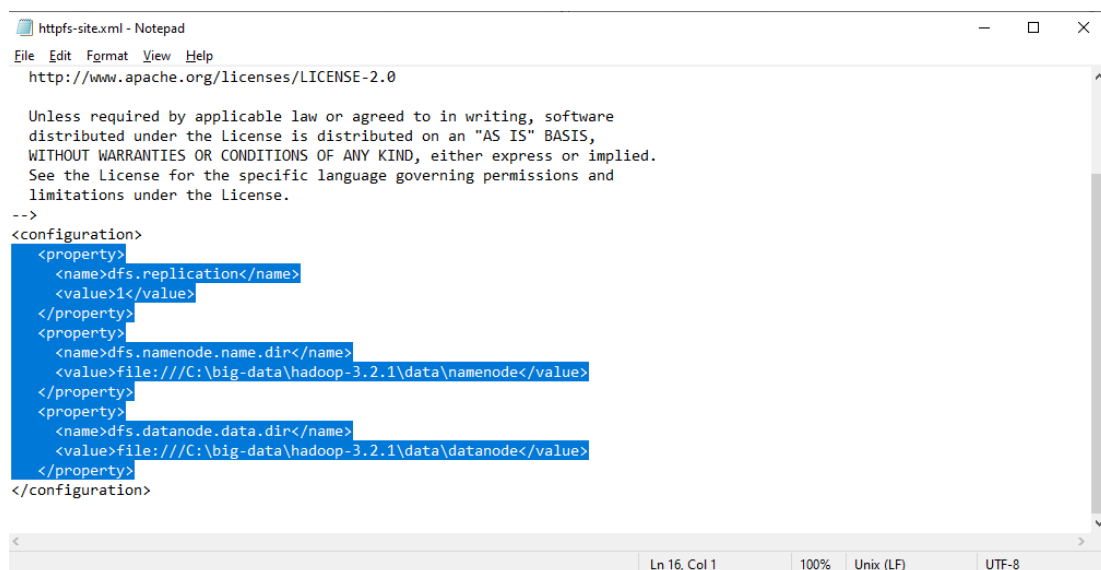
Now, setup Hadoop. Open your hadoop folder, open etc, then hadoop. Edit the xml file named core-site. Enter the following highlighted on the picture.

*core-site.xml - Notepad

File   Edit   Format   View   Help

```xml
<!--
  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://0.0.0.0:19000</value>
    </property>
</configuration>
```

Now, edit another xml file. It is named https-site. Put the highlighted code below. Take note that the directory on the <value> differs on your end. Make sure that you created a folder on you Hadoop folder named "data". On the data folder, create two folders named "namenode" and "datanode".

httpfs-site.xml - Notepad

File   Edit   Format   View   Help

```xml
  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.
-->
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>file:///C:\big-data\hadoop-3.2.1\data\namenode</value>
    </property>
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>file:///C:\big-data\hadoop-3.2.1\data\datanode</value>
    </property>
</configuration>
```

Ln 16, Col 1   100%   Unix (LF)   UTF-8

Edit another xml, mapred-site.

mapred-site.xml - Notepad

File   Edit   Format   View   Help

```xml
<?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.application.classpath</name>
        <value>%HADOOP_HOME%/share/hadoop/mapreduce/*,%HADOOP_HOME%/share/hadoop/mapreduce/lib/*,%HADOOP_HOME%/share/hadoop/common/*,%HADOOP_
    </property>
</configuration>
```
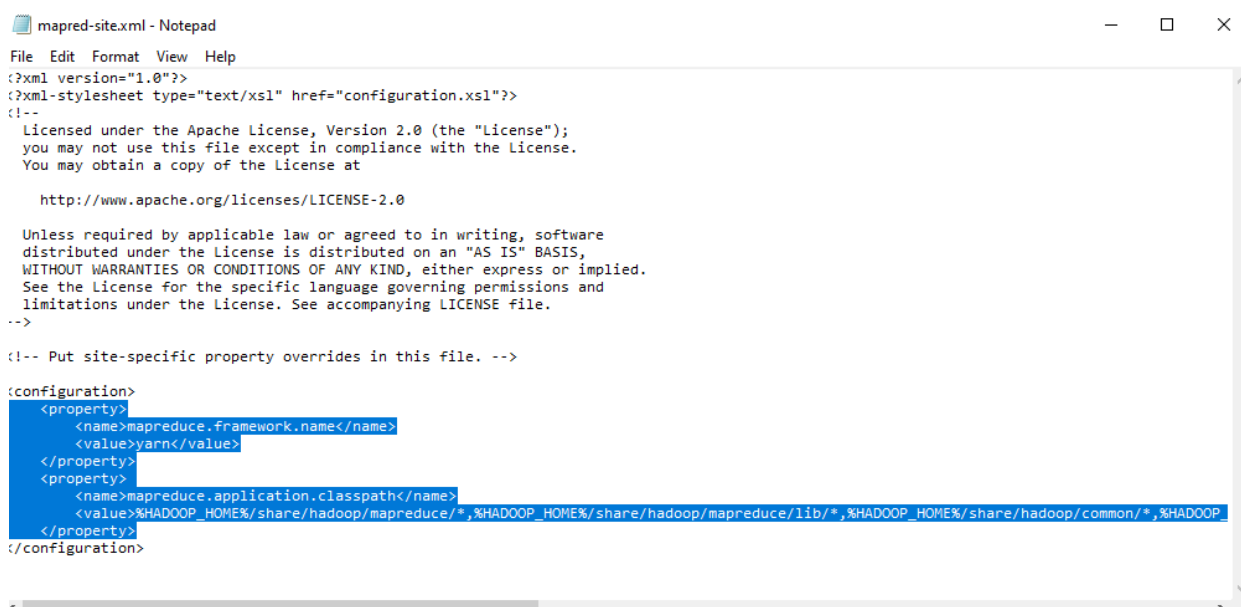
Edit last xml, yarn-site.

Now, on your cmd, initiate this command: hdfs namenode -format.

Initiate hdfs daemons. Open new cmd and type "%HADOOP_HOME%\sbin\start-dfs.cmd" on your cmd.

Initiate yarn daemons. Open cmd (run as administrator) and type this command "%HADOOP_HOME%\sbin\start-yarn.cmd".

On your browser, open this link that will show HDFS Namenode information UI: http://localhost:9870/dfshealth.html#tab-overview. This site should appear as shown below.

```
Apache Hadoop Distribution - yarn  nodemanager                                    —    □    ×
INFO: Registering org.apache.hadoop.yarn.server.nodemanager.webapp.JAXBContextResolver as a provider class
Sep 26, 2023 9:26:11 PM com.sun.jersey.server.impl.application.WebApplicationImpl _initiate
INFO: Initiating Jersey application, version 'Jersey: 1.19 02/11/2015 03:25 AM'
Sep 26, 2023 9:26:11 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.server.nodemanager.webapp.JAXBContextResolver to GuiceManagedComponentProvider with
 the scope "Singleton"
Sep 26, 2023 9:26:11 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.webapp.GenericExceptionHandler to GuiceManagedComponentProvider with the scope "Sin
gleton"
Sep 26, 2023 9:26:12 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.server.nodemanager.webapp.NMWebServices to GuiceManagedComponentProvider with the s
cope "Singleton"
2023-09-26 21:26:12,325 INFO handler.ContextHandler: Started o.e.j.w.WebAppContext@61a1ea2c{/,file:///C:/Users/admin/App
Data/Local/Temp/jetty-0.0.0.0-8042-node-_-any-4016689139982644232.dir/webapp/,AVAILABLE}{/node}
2023-09-26 21:26:12,342 INFO server.AbstractConnector: Started ServerConnector@63411512{HTTP/1.1,[http/1.1]}{0.0.0.0:804
2}
2023-09-26 21:26:12,343 INFO server.Server: Started @28164ms
2023-09-26 21:26:12,345 INFO webapp.WebApps: Web app node started at 8042
2023-09-26 21:26:12,347 INFO nodemanager.NodeStatusUpdaterImpl: Node ID assigned is : DESKTOP-QIL2EDF:60838
2023-09-26 21:26:12,350 INFO util.JvmPauseMonitor: Starting JVM pause monitor
2023-09-26 21:26:12,376 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8031
2023-09-26 21:26:12,458 INFO nodemanager.NodeStatusUpdaterImpl: Sending out 0 NM container statuses: []
2023-09-26 21:26:12,485 INFO nodemanager.NodeStatusUpdaterImpl: Registering with RM using containers :[]
2023-09-26 21:26:12,925 INFO security.NMContainerTokenSecretManager: Rolling master-key for container-tokens, got key wi
th id 351281418
2023-09-26 21:26:12,927 INFO security.NMTokenSecretManagerInNM: Rolling master-key for container-tokens, got key with id
 1538005558
2023-09-26 21:26:12,929 INFO nodemanager.NodeStatusUpdaterImpl: Registered with ResourceManager as DESKTOP-QIL2EDF:60838
 with total resource of <memory:8192, vCores:8>
```


On your browser again, open this another link that will show HDFS Datanode information UI:
http://localhost:9864/datanode.html.

Open this link on another tab that will direct you to YARN resource manager UI: http://localhost:8088

It shows that Hadoop is now working on your end.

HIVE INSTALLATION

In google, we will search apache hive 2.1.0 download. Click on the first link that appears like the one below.

We will click on apache-hive-2.1.0-bin.tar.gz file. This will begin the download.

We will extract the apache-hive-2.1.0-bin folder to the C drive. Then we will rename the folder to **hive** so that we can access it easily later.
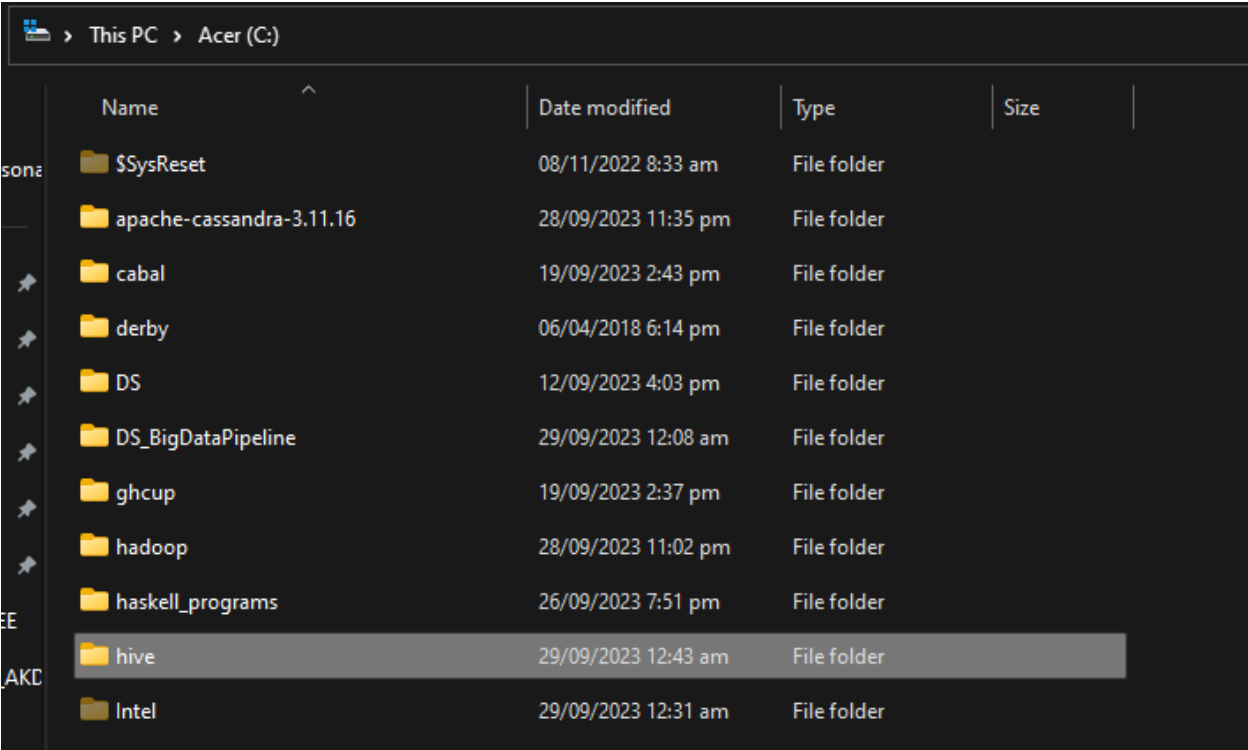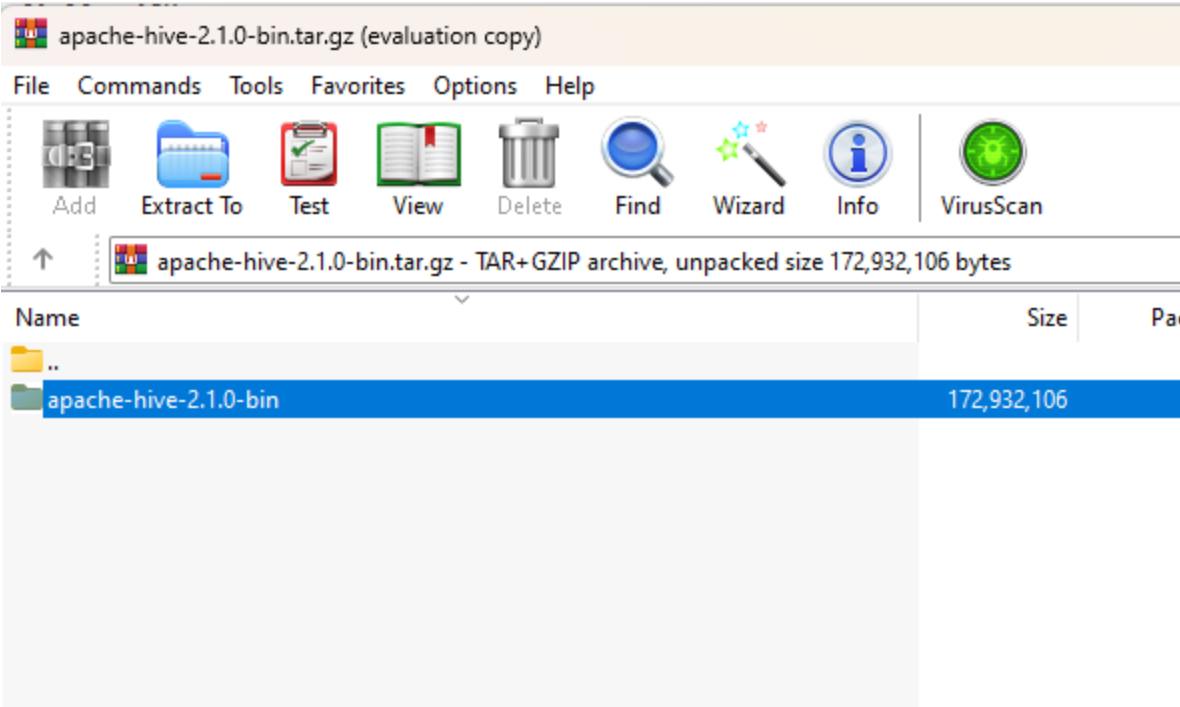
DERBY INSTALLATION

Now in google we will search apache derby 10.12.1.1 download. Click on the first link that shows like the one below.
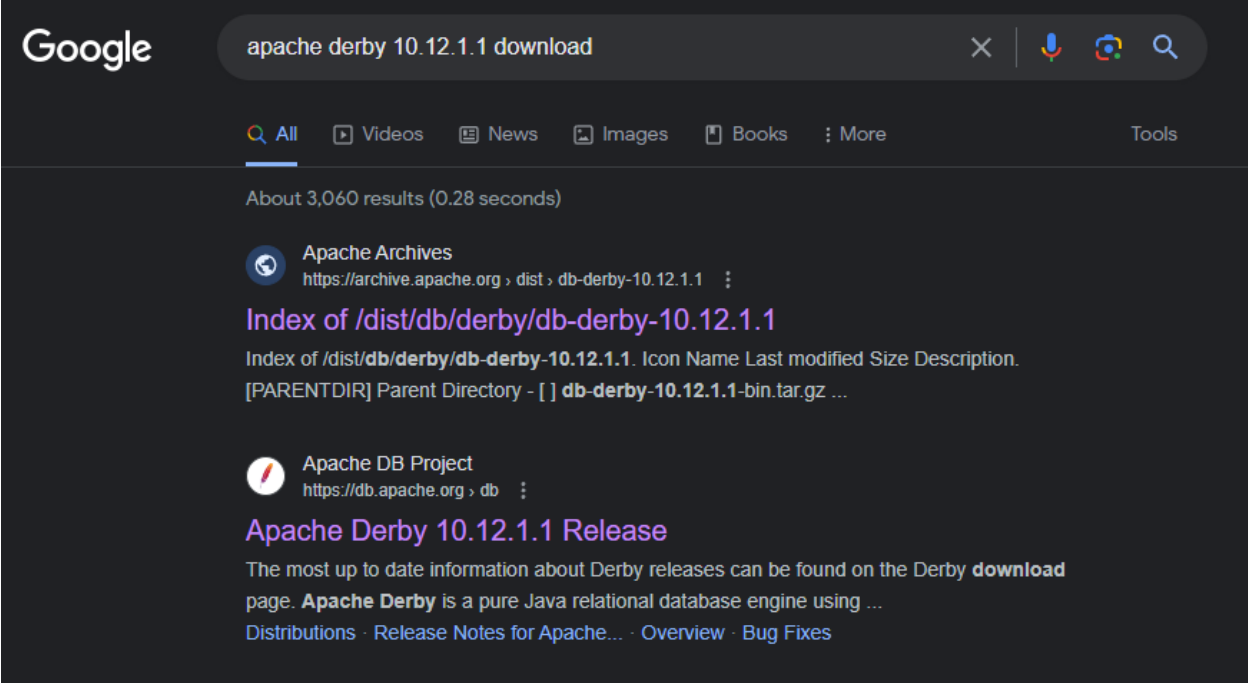
We will click on db-derby-10.12.1.1-bin.tar.gz file. This will begin the download of the file.

Index of /dist/db/derby/db-derby-10.12.1.1

Name	Last modified	Size	Description
Parent Directory		-	
db-derby-10.12.1.1-bin.tar.gz	2015-10-10 14:38	18M	
db-derby-10.12.1.1-bin.tar.gz.asc	2015-10-10 14:38	194	
db-derby-10.12.1.1-bin.tar.gz.md5	2015-10-10 14:38	33	
db-derby-10.12.1.1-bin.zip	2015-10-10 14:38	20M	

Again we will extract the file to the C drive. Then we will rename it to derby so we can easily access the folder later.

db-derby-10.12.1.1-bin.tar.gz (evaluation copy)

File   Commands   Tools   Favorites   Options   Help

Add   Extract To   Test   View   Delete   Find   Wizard   Info   VirusScan

db-derby-10.12.1.1-bin.tar.gz - TAR+GZIP archive, unpacked size 35,397,502 bytes

Name	Size	Packed	Type
..			File folder
db-derby-10.12.1.1-bin	35,397,502	?	File folder

This PC › Acer (C:) ›

Name	Date modified	Type
$SysReset	08/11/2022 8:33 am	File folder
apache-cassandra-3.11.16	28/09/2023 11:35 pm	File folder
cabal	19/09/2023 2:43 pm	File folder
derby	29/09/2023 12:49 am	File folder
DS	12/09/2023 4:03 pm	File folder

We need the hive-site.xml to configure hive. Download the xml file from the link below.

https://drive.google.com/file/d/1qqAo7RQfr5Q6O-GTom6Rji3TdufP81zd/view

After the download, move the hive-site.xml file to hive conf folder.

This PC › Acer (C:) › hive › conf

Name	Date modified	Type	Size
beeline-log4j2.properties	03/06/2016 6:43 pm	Dev-C++ Templat...	2 KB
hive-default.xml	17/06/2016 8:03 am	Dev-C++ Templat...	221 KB
hive-env.sh	03/06/2016 6:43 pm	Dev-C++ Templat...	3 KB
hive-exec-log4j2.properties	03/06/2016 6:43 pm	Dev-C++ Templat...	3 KB
hive-log4j2.properties	03/06/2016 6:43 pm	Dev-C++ Templat...	3 KB
ivysettings	10/06/2016 5:00 pm	XML File	3 KB
llap-cli-log4j2.properties	03/06/2016 6:43 pm	Dev-C++ Templat...	3 KB
llap-daemon-log4j2.properties	03/06/2016 6:43 pm	Dev-C++ Templat...	5 KB
parquet-logging	09/06/2016 2:47 am	PROPERTIES File	3 KB
hive-site	29/09/2023 12:51 am	XML File	2 KB

Then we will go to the derby directory. Inside it we will go to the lib directory. We will select all jar files in the directory and copy them. Then we will paste them in the hive lib directory as shown below.
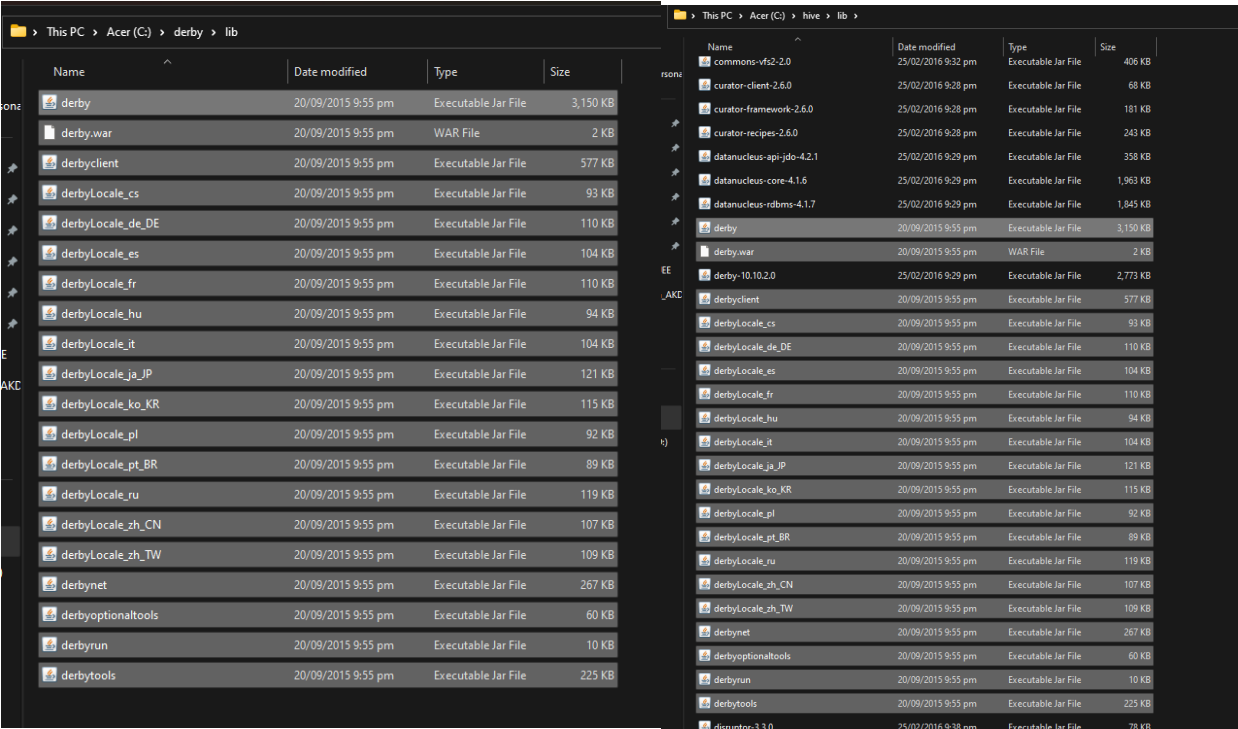
Now we will set the paths in the environment variables. First set **HIVE_HOME** and set the value to the path of the hive directory.

Next set **HIVE_BIN**. Set the value to the path of the bin directory.

Next set **HIVE_LIB.** Set the value to the path of the lib directory.

Now we will set **DERBY_HOME**. Enter the path of the derby directory.

Next set **HADOOP_USER_CLASSPATH_FIRST**. The value is set to true.

Edit User Variable

Variable name: HADOOP_USER_CLASSPATH_FIRST

Variable value: true

Browse Directory... Browse File... OK Cancel

Lastly, we will add **C:\hive\bin** and **C:\derby\bin** to the systems path. Make sure to click **OK** to save all changes.

Now we will run Hadoop. We will run a command prompt in administrator mode. Then we will access the sbin directory of Hadoop.

We will run the nodes by entering the command **start-dfs**. We will also run yarn by entering the command **start-yarn**.
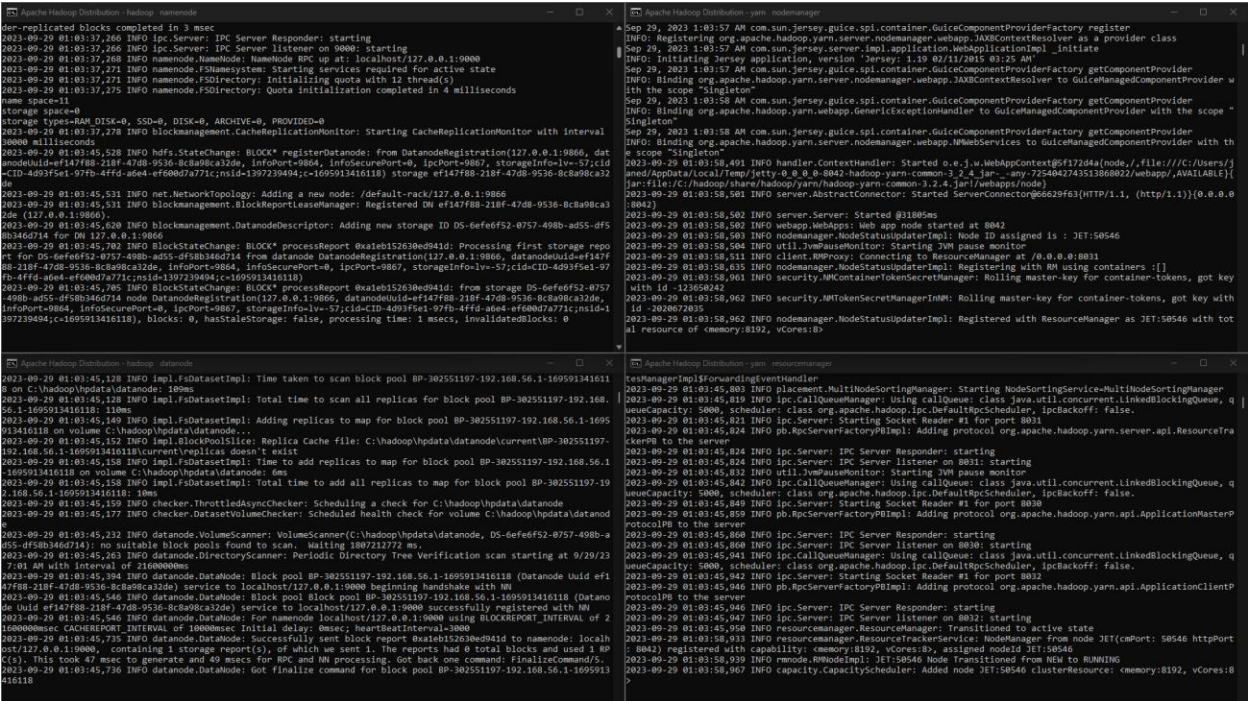
Next, we will run the derby server. In a new administrator command prompt, we will execute the command **StartNetworkServer -h 0.0.0.0** as this will start the server.
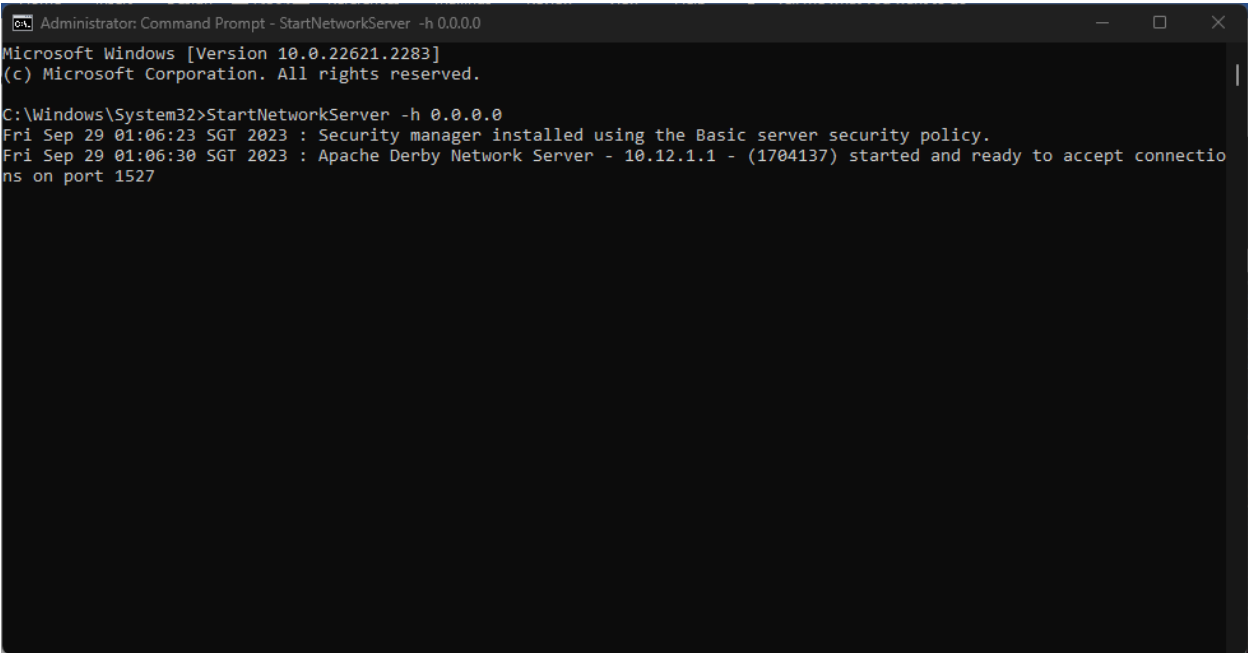
Now, we can run hive. We will open a new administrator command prompt and navigate to the bin directory of hive. Then we will enter **hive**.
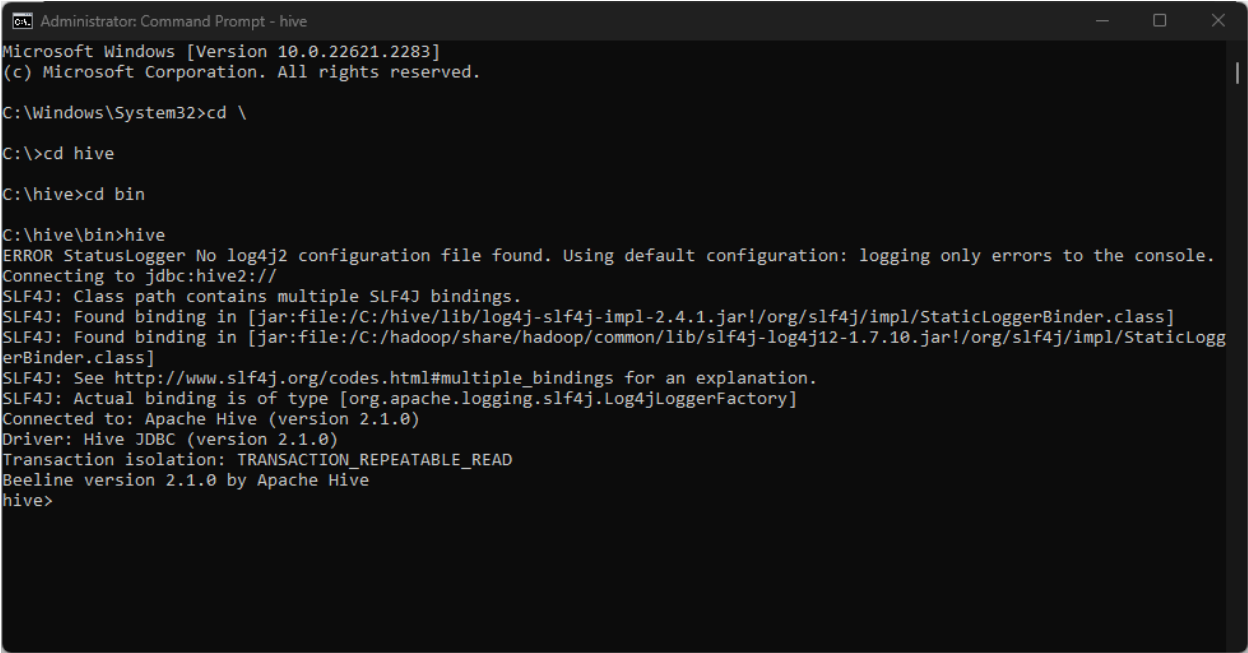
To prepare for migration, we will create a database. Inside the database we will create a table that will store the data from cassandra. We will run the command **CREATE DATABASE cctvdata;** for creating the database and **SHOW DATABASES;** to verify that it was created.

```
C:\hive\bin>hive
ERROR StatusLogger No log4j2 configuration file found. Using default configuration: logging only errors to the console.
Connecting to jdbc:hive2://
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/hive/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/C:/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLogg
erBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Connected to: Apache Hive (version 2.1.0)
Driver: Hive JDBC (version 2.1.0)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 2.1.0 by Apache Hive
hive> CREATE DATABASE cctvdata;
OK
No rows affected (0.866 seconds)
hive> SHOW DATABASES;
OK
cctvdata
default
test
3 rows selected (0.212 seconds)
hive>
```

We will create a table inside the cctvdata database. The code we run was **CREATE TABLE count_cctv(timeuuid_id STRING, bike INT, bus INT, car INT, date_saved DATE, jeepney INT, lgu_code STRING, other INT, sensor_id STRING, time_saved TIMESTAMP, total INT, truck INT, tryke INT);** Then we verified that the table was created correctly be entering **DESCRIBE count_cctv;**


```
hive> SHOW DATABASES;
OK
cctvdata
default
test
3 rows selected (0.212 seconds)
hive> use cctvdata;
OK
No rows affected (0.027 seconds)
hive> CREATE TABLE count_cctv(timeuuid_id STRING, bike INT, bus INT, car INT, date_saved DATE, jeepney INT, lgu_code STR
ING, other INT, sensor_id STRING, time_saved TIMESTAMP, total INT, truck INT, tryke INT);
OK
No rows affected (0.898 seconds)
hive> DESCRIBE count_cctv;
OK
timeuuid_id string
bike int
bus int
car int
date_saved date
jeepney int
lgu_code string
other int
sensor_id string
time_saved timestamp
total int
truck int
tryke int
13 rows selected (0.441 seconds)
hive>
```

CASSANDRA TO HIVE MIGRATION

We created a python program using pyspark to migrate data from cassandra to hive. In the program, we import SparkSession so that we can use pyspark functions and date, datetime, and timedelta to handle data and time functionailies. In the code, we first specified the cassandra configuration which include the host, port, target keyspace and table. Then we specified the hive configuration which were the database and table where the data will be migrated to. Then a spark session is created to handle the data migration. The date of the day and timestamp for the end of the day, which is 11:59 are initialized. The spark session will then read through the cassandra database. Afterwards it will filter through the data and will get all data that stored inside the database within the day only. The data is sent to a temporary dataframe. Afterwards all the data is inserted into the hive count_cctv table inside the cctvdata database.

```python
from pyspark.sql import SparkSession
from pyspark.sql.functions import current_date
from datetime import date, datetime, timedelta

# Cassandra Configuration
cassandra_host = "127.0.0.1"
cassandra_port = "9042"
cassandra_keyspace = "keyspacedatasci"
cassandra_table = "counts_cctv"

# Hive Configuration
hive_database = "cctvdata"
hive_table = "count_cctv"

# Initialize Spark session
spark = SparkSession.builder \
    .appName("CassandraToHiveMigration") \
    .config("hive.metastore.uris", "thrift://0.0.0.0:9083") \
    .master("local[*]") \
    .config("spark.cassandra.connection.host", cassandra_host) \
    .config("spark.cassandra.connection.port", cassandra_port) \
    .config("spark.jars", "C:\\DS_BigDataPipeline\\lib\\kafka-clients-3.5.1.jar,"
                          "C:\\DS_BigDataPipeline\\lib\\spark-streaming-kafka-0-10_2.12-3.0.3.jar,"
                          "C:\\DS_BigDataPipeline\\lib\\spark-cassandra-connector_2.12-3.0.1.jar") \
    .config("spark.jars.packages", "org.apache.spark:spark-sql-kafka-0-10_2.12:3.0.3,"
                                    "com.datastax.spark:spark-cassandra-connector_2.12:3.0.1") \
    .enableHiveSupport() \
    .getOrCreate()
```

```
29
30   # Get today's date
31   today_date = date.today()
32
33   # Calculate the end of the day timestamp (11:59:59 PM)
34   end_of_day_timestamp = datetime.combine(today_date, datetime.min.time()) + timedelta(hours=23, minutes=59, seconds=59)
35
36   # Read data from Cassandra
37   cassandra_df = spark.read \
38       .format("org.apache.spark.sql.cassandra") \
39       .options(keyspace=cassandra_keyspace, table=cassandra_table) \
40       .load()
41
42   # Filter data for today up to the end of the day
43   filtered_df = cassandra_df.filter(
44       (cassandra_df.date_saved.cast("timestamp") >= datetime.combine(today_date, datetime.min.time())) &
45       (cassandra_df.date_saved.cast("timestamp") <= end_of_day_timestamp)
46   )
47
48
49   # Register the DataFrame as a temporary table
50   filtered_df.createOrReplaceTempView("cctv_data")
51
52   # Insert data into Hive
53   spark.sql("INSERT INTO cctvdata.count_cctv SELECT * FROM cctv_data")
54
55   # Stop the Spark session
56   spark.stop()
57
```

We run the code in the administrator command prompt. We enter the command **python hive_migration.py**. Then the program executes and carries out the data migration from cassandra to hive.

```
Administrator: Command Prompt                                          —  □  ✕

(BDvenv) C:\DS_BigDataPipeline>python hive_migration.py
Warning: Ignoring non-Spark config property: hive.metastore.uris
Ivy Default Cache set to: C:\Users\janed\.ivy2\cache
The jars for the packages stored in: C:\Users\janed\.ivy2\jars
:: loading settings :: url = jar:file:/C:/Spark/spark-3.0.3-bin-hadoop3.2/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settin
gs/ivysettings.xml
org.apache.spark#spark-sql-kafka-0-10_2.12 added as a dependency
com.datastax.spark#spark-cassandra-connector_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-3f8bf576-ef91-4849-afef-7e3a910a3e2c;1.0
        confs: [default]
        found org.apache.spark#spark-sql-kafka-0-10_2.12;3.0.3 in central
        found org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.0.3 in central
        found org.apache.kafka#kafka-clients;2.4.1 in central
        found com.github.luben#zstd-jni;1.4.4-3 in central
        found org.lz4#lz4-java;1.7.1 in central
        found org.xerial.snappy#snappy-java;1.1.8.2 in central
        found org.slf4j#slf4j-api;1.7.30 in central
        found org.spark-project.spark#unused;1.0.0 in central
        found org.apache.commons#commons-pool2;2.6.2 in central
        found com.datastax.spark#spark-cassandra-connector_2.12;3.0.1 in central
        found com.datastax.spark#spark-cassandra-connector-driver_2.12;3.0.1 in central
        found com.datastax.oss#java-driver-core-shaded;4.10.0 in central
        found com.datastax.oss#native-protocol;1.4.12 in central
        found com.datastax.oss#java-driver-shaded-guava;25.1-jre-graal-sub-1 in central
        found com.typesafe#config;1.4.1 in central
        found io.dropwizard.metrics#metrics-core;4.1.16 in central
        found org.hdrhistogram#HdrHistogram;2.1.12 in central
        found org.reactivestreams#reactive-streams;1.0.3 in central
        found com.github.stephenc.jcip#jcip-annotations;1.0-1 in central
```

```
Administrator: Command Prompt - python hive_migration.py                —  □  ✕

        org.apache.commons#commons-pool2;2.6.2 from central in [default]
        org.apache.kafka#kafka-clients;2.4.1 from central in [default]
        org.apache.spark#spark-sql-kafka-0-10_2.12;3.0.3 from central in [default]
        org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.0.3 from central in [default]
        org.hdrhistogram#HdrHistogram;2.1.12 from central in [default]
        org.lz4#lz4-java;1.7.1 from central in [default]
        org.reactivestreams#reactive-streams;1.0.3 from central in [default]
        org.scala-lang#scala-reflect;2.12.11 from central in [default]
        org.slf4j#slf4j-api;1.7.30 from central in [default]
        org.spark-project.spark#unused;1.0.0 from central in [default]
        org.xerial.snappy#snappy-java;1.1.8.2 from central in [default]
        :: evicted modules:
        org.slf4j#slf4j-api;1.7.26 by [org.slf4j#slf4j-api;1.7.30] in [default]
        ---------------------------------------------------------------------
        |                    |          modules         ||    artifacts    |
        |        conf        | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default       |  27   |  0   |  0   |  1   ||  26  |  0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-3f8bf576-ef91-4849-afef-7e3a910a3e2c
        confs: [default]
        0 artifacts copied, 26 already retrieved (0kB/29ms)
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
C:\DS_BigDataPipeline\BDvenv\lib\site-packages\pyspark\context.py:227: DeprecationWarning: Support for Python 2 and Pyth
on 3 prior to version 3.6 is deprecated as of Spark 3.0. See also the plan for dropping Python 2 support at https://spar
k.apache.org/news/plan-for-dropping-python-2-support.html.
  DeprecationWarning)
```

After executing the program, we can verify if the migration is complete by checking the hive database and table. We used the command **SELECT * FROM count_cctv;** to show all rows in the table.

```
hive> SELECT * FROM COUNT_CCTV;
-chgrp: 'JET\janed' does not match expected pattern for group
Usage: hadoop fs [generic options] -chgrp [-R] GROUP PATH...
OK
1b38f180-5e2d-11ee-b563-089798cc0a5d 2 1 4 2023-09-29 2 1260 1 sensor_07  13 1 2
e349070f-5e2c-11ee-8c76-089798cc0a5d 1 2 4 2023-09-29 2 1220 1 sensor_03  15 2 3
f2567b70-5e2c-11ee-9215-089798cc0a5d 0 0 2 2023-09-29 1 1200 1 sensor_01  9 2 3
3a67bb40-5e2d-11ee-9d1f-089798cc0a5d 1 0 1 2023-09-29 0 1250 1 sensor_06  7 1 3
0374ce70-5e2d-11ee-859f-089798cc0a5d 0 0 3 2023-09-29 0 1240 2 sensor_05  8 0 3
3444afc0-5e2d-11ee-8641-089798cc0a5d 5 2 2 2023-09-29 0 1240 2 sensor_05  14 1 2
443446c0-5e2d-11ee-b627-089798cc0a5d 1 1 0 2023-09-29 2 1260 1 sensor_07  8 1 2
3c37e761-5e2d-11ee-95af-089798cc0a5d 3 2 4 2023-09-29 0 1250 0 sensor_06  11 2 0
13b7dca1-5e2d-11ee-a4b8-089798cc0a5d 3 1 1 2023-09-29 1 1240 2 sensor_05  8 0 0
28bc4370-5e2d-11ee-9b40-089798cc0a5d 2 0 4 2023-09-29 1 1290 1 sensor_10  8 0 0
0dfc6c8f-5e2d-11ee-87d4-089798cc0a5d 3 1 1 2023-09-29 2 1280 0 sensor_09  10 2 1
e296a700-5e2c-11ee-ab09-089798cc0a5d 0 2 3 2023-09-29 2 1210 1 sensor_02  10 0 2
e77ec180-5e2c-11ee-83d8-089798cc0a5d 1 2 1 2023-09-29 1 1200 2 sensor_01  11 1 3
de78ba4f-5e2c-11ee-a87b-089798cc0a5d 3 1 2 2023-09-29 1 1220 0 sensor_03  7 0 0
0eddcbde-5e2d-11ee-b6f9-089798cc0a5d 4 2 1 2023-09-29 1 1230 0 sensor_04  12 1 2
ec5799c0-5e2d-11ee-b8ea-089798cc0a5d 5 2 3 2023-09-29 1 1270 1 sensor_08  12 0 0
257c5ab0-5e2d-11ee-8e80-089798cc0a5d 1 0 2 2023-09-29 2 1280 2 sensor_09  7 0 0
e98a7000-5e2d-11ee-915d-089798cc0a5d 5 1 1 2023-09-29 0 1240 2 sensor_05  9 0 0
e8197ae1-5e2c-11ee-986d-089798cc0a5d 5 0 2 2023-09-29 2 1200 0 sensor_01  9 0 0
1e4bafc0-5e2d-11ee-8019-089798cc0a5d 4 2 1 2023-09-29 1 1220 1 sensor_03  10 1 0
f2b58b5e-5e2c-11ee-a461-089798cc0a5d 2 2 3 2023-09-29 1 1250 2 sensor_06  15 2 3
62630 5a1-5e2d-11ee-87a3-089798cc0a5d 1 1 0 2023-09-29 1 1270 1 sensor_08  7 2 1
fd183c61-5e2c-11ee-9a03-089798cc0a5d 0 2 0 2023-09-29 0 1260 1 sensor_07  7 2 2
24efd221-5e2d-11ee-b0a2-089798cc0a5d 4 2 4 2023-09-29 2 1200 2 sensor_01  19 2 3
ceacd7f0-5e2c-11ee-8589-089798cc0a5d 2 1 3 2023-09-29 2 1260 2 sensor_07  10 0 2
ecd41cc0-5e2c-11ee-af63-089798cc0a5d 1 0 3 2023-09-29 2 1260 1 sensor_07  8 1 0
fccc1830-5e2c-11ee-b09d-089798cc0a5d 0 1 3 2023-09-29 0 1230 0 sensor_04  6 0 2
ea1e2480-5e2c-11ee-a36b-089798cc0a5d 4 1 1 2023-09-29 1 1230 2 sensor_04  10 0 1
1c235180-5e2d-11ee-9174-089798cc0a5d 4 0 2 2023-09-29 2 1240 2 sensor_05  11 1 0
28ca9b4f-5e2d-11ee-a5b7-089798cc0a5d 5 0 1 2023-09-29 1 1230 0 sensor_04  9 0 2
dfd7ae11-5e2c-11ee-b354-089798cc0a5d 3 1 1 2023-09-29 0 1290 2 sensor_10  11 2 2
14b0e2a1-5e2d-11ee-be70-089798cc0a5d 4 2 1 2023-09-29 0 1270 1 sensor_08  11 2 1
55abdd4f-5e2d-11ee-90d6-089798cc0a5d 5 1 4 2023-09-29 0 1260 0 sensor_07  12 1 1
271ee71e-5e2d-11ee-8e4d-089798cc0a5d 0 2 1 2023-09-29 0 1250 1 sensor_06  5 1 0
221ed291-5c92-11ee-a71b-dc41a9636cc0 4 0 4 2023-09-27 1 1200 1 sensor_01  14 1 3
0712b5b0-5e2d-11ee-a1c4-089798cc0a5d 1 1 0 2023-09-29 1 1240 1 sensor_05  7 1 3
f4ab40e1-5e2c-11ee-8963-089798cc0a5d 5 2 1 2023-09-29 2 1210 1 sensor_02  12 0 1
d1f85970-5e2d-11ee-917c-089798cc0a5d 5 2 1 2023-09-29 0 1290 0 sensor_10  11 0 3
0d5d1f51-5e2d-11ee-9768-089798cc0a5d 3 1 4 2023-09-29 1 1260 0 sensor_07  11 2 0
cf053121-5e2c-11ee-be0e-089798cc0a5d 2 1 4 2023-09-29 1 1240 1 sensor_05  14 2 3
eef9bbe1-5e2c-11ee-9991-089798cc0a5d 3 1 3 2023-09-29 2 1240 1 sensor_05  11 1 0
179a6e00-5e2d-11ee-817b-089798cc0a5d 3 0 2 2023-09-29 2 1210 2 sensor_02  12 2 1
5790c9f0-5e2d-11ee-8b6c-089798cc0a5d 2 2 1 2023-09-29 2 1220 0 sensor_03  8 0 1
319b3a9e-5e2d-11ee-b110-089798cc0a5d 3 0 4 2023-09-29 1 1200 0 sensor_01  8 0 0
ea745acf-5e2c-11ee-ae62-089798cc0a5d 3 1 2 2023-09-29 1 1280 0 sensor_09  9 0 2
0f754bae-5c92-11ee-830f-dc41a9636cc0 2 1 4 2023-09-27 0 1200 1 sensor_01  9 1 0
e0e54d80-5e2c-11ee-9a4b-089798cc0a5d 4 2 2 2023-09-29 2 1270 2 sensor_08  15 0 3
689bbc00-5e2d-11ee-b305-089798cc0a5d 5 0 4 2023-09-29 1 1290 1 sensor_10  11 0 0
568b40d1-5e2d-11ee-8e08-089798cc0a5d 4 1 2 2023-09-29 1 1290 0 sensor_10  10 1 1
d54fc1cf-5e2c-11ee-99c7-089798cc0a5d 4 2 3 2023-09-29 2 1210 1 sensor_02  16 2 2
2e83734f-5e2d-11ee-a15b-089798cc0a5d 0 0 2 2023-09-29 0 1260 1 sensor_07  4 1 0
3746a521-5e2c-11ee-82d1-089798cc0a5d 0 2 3 2023-09-29 0 1240 1 sensor_05  10 2 2
01fe0cf0-5e2d-11ee-80ab-089798cc0a5d 5 2 4 2023-09-29 1 1200 1 sensor_01  16 1 2
53dce9b0-5e2d-11ee-a46f-089798cc0a5d 0 1 1 2023-09-29 1 1280 2 sensor_09  6 0 1
d117965e-5e2c-11ee-9abf-089798cc0a5d 5 1 1 2023-09-29 0 1210 2 sensor_02  11 0 2
d9639c5e-5e2c-11ee-81fa-089798cc0a5d 0 0 2 2023-09-29 2 1250 2 sensor_06  9 0 3
63871340-5e2d-11ee-9037-089798cc0a5d 5 1 4 2023-09-29 2 1280 0 sensor_09  17 2 3
279612f0-5e2d-11ee-adcf-089798cc0a5d 4 1 0 2023-09-29 0 1230 2 sensor_04  11 1 3
56f637a1-5e2d-11ee-b8c7-089798cc0a5d 5 2 1 2023-09-29 0 1200 1 sensor_01  13 2 2
da412b21-5e2c-11ee-9ffe-089798cc0a5d 1 0 1 2023-09-29 2 1200 2 sensor_01  8 0 2
6a9c9511-5e2d-11ee-8df1-089798cc0a5d 5 0 4 2023-09-29 2 1250 1 sensor_06  13 1 0
56e0dade-5e2d-11ee-8e43-089798cc0a5d 2 0 2 2023-09-29 1 1280 0 sensor_09  9 1 3
d80f7e11-5e2c-11ee-a01b-089798cc0a5d 2 0 1 2023-09-29 0 1240 1 sensor_05  6 2 0
39ccb3c0-5e2d-11ee-85fa-089798cc0a5d 5 0 0 2023-09-29 2 1250 1 sensor_06  9 1 0
686c6ea1-5e2d-11ee-b727-089798cc0a5d 0 1 4 2023-09-29 1 1270 0 sensor_08  9 0 3
09d09d30-5e2d-11ee-b7e1-089798cc0a5d 0 0 2 2023-09-29 1 1220 1 sensor_03  4 0 0
51f5b31e-5e2d-11ee-b7bf-089798cc0a5d 3 2 3 2023-09-29 1 1230 2 sensor_04  11 0 0
0f81fb1e-5e2d-11ee-b997-089798cc0a5d 4 1 0 2023-09-29 1 1240 0 sensor_05  8 2 0
dcdda7ee-5e2c-11ee-9284-089798cc0a5d 0 0 1 2023-09-29 2 1230 1 sensor_04  4 0 0
02851740-5e2d-11ee-bf46-089798cc0a5d 5 1 1 2023-09-29 1 1210 1 sensor_02  12 2 1
40ac5880-5e2d-11ee-9ee9-089798cc0a5d 0 1 1 2023-09-29 1 1200 1 sensor_01  7 1 2
11a884a1-5e2d-11ee-aec4-089798cc0a5d 2 2 1 2023-09-29 2 1200 2 sensor_01  14 2 3
2,755 rows selected (1.032 seconds)
hive>
```

References:

Raymond (2021). *Apache Hive 3.0.0 Installation on Windows 10 Step by Step Guide*. Kontext. https://kontext.tech/article/291/apache-hive-300-installation-on-windows-10-step-by-step-guide

Raymond (2021). *Install Hadoop 3.2.1 on Windows 10 Step by Step Guide*. Kontext. https://kontext.tech/article/377/latest-hadoop-321-installation-on-windows-10-step-by-step-guide

Kulkarni, M. (2018). *EASY HIVE INSTALLATION ON WINDOWS*. YouTube. https://www.youtube.com/watch?v=npyRXkMhrgk

OnlineLearningCenter (2016). *How To Install Hive In Hadoop On Ubuntu | Hive | Hadoop Tutorial |@OnlineLearningCenterIndia*. YouTube. https://www.youtube.com/watch?v=D_rHkYDKSeE

IvyProSchool (2023). *How to Install Hadoop in Windows 10 & 11 | Data Engineering Tutorials | IvyProSchool*. YouTube. https://www.youtube.com/watch?v=knAS0w-jiUk&t=923s

Fadlallah, H. (2020). *Installing Apache Hive 3.1.2 on Windows 10*. Medium. https://towardsdatascience.com/installing-apache-hive-3-1-2-on-windows-10-70669ce79c79