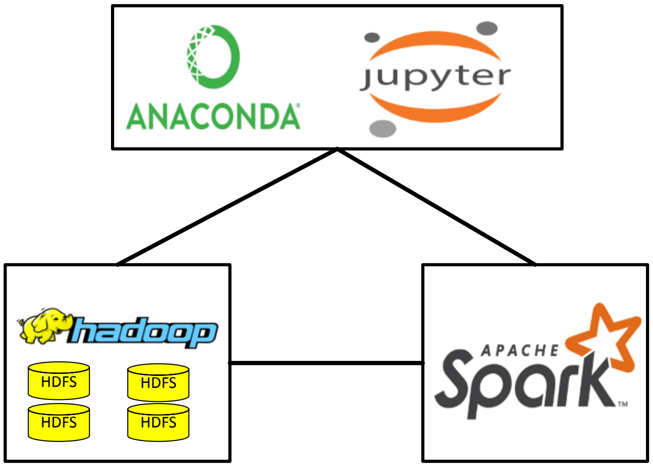
Build Machine learning Lab

Data science Lab

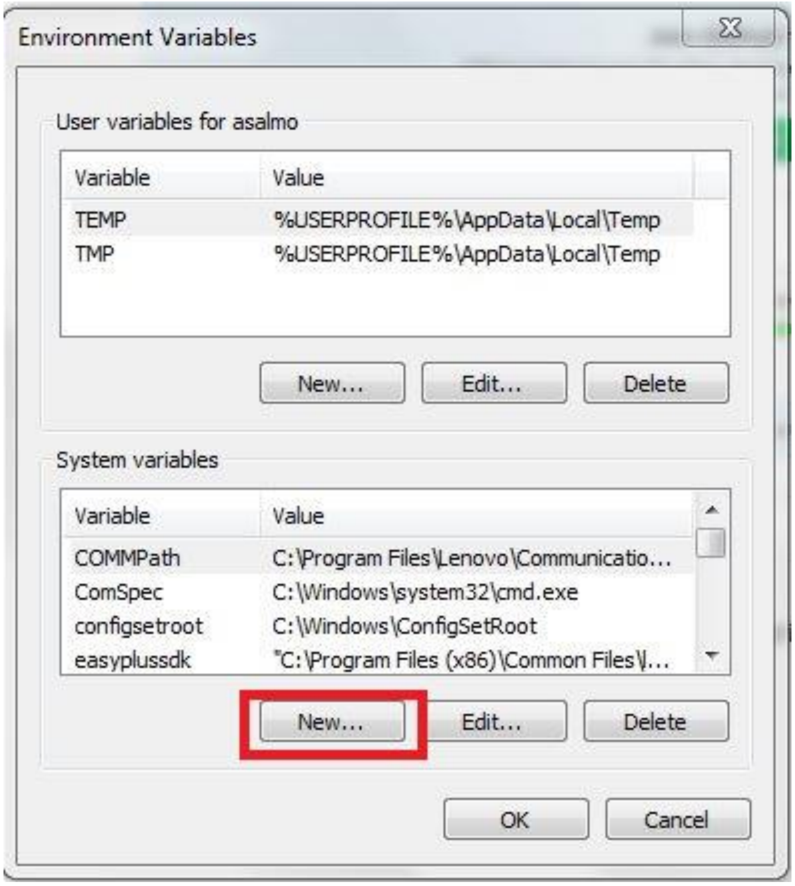


- Java (OpenJDK 8)
- Hadoop (using version: 3.1.2)
- Spark (using version: 2.7.0 to 3.2.0 preview)
- Anaconda (Jupyter). Windows 7: 2018 and Windows 10: 2019

1- If you don't have Java 8, you will need to implement step 1 & 2. If you already have Java 8, you can pass step # 1. Hadoop 3.1.2 works with Java 8.

Install OpenJDK-8 Java
Download and install
(https://developers.redhat.com/products/openjdk/download?extIdCarryOver=true&sc_cid=701f2000001OH7JAAW)

2- Add JAVA_HOME to environment variables.



Environment Variables

User variables for asalmo

Variable	Value
TEMP	%USERPROFILE%\AppData\Local\Temp
TMP	%USERPROFILE%\AppData\Local\Temp

System variables

Variable	Value
COMMPath	C:\Program Files\Lenovo\Communicatio...
ComSpec	C:\Windows\system32\cmd.exe
configsetroot	C:\Windows\ConfigSetRoot
easyplussdk	"C:\Program Files (x86)\Common Files\...

JAVA_HOME= C:\alaa\AdoptOpenJDK\jdk-8.0.232.09-hotspot
Add JAVA_HOME to path %JAVA_HOME%\bin

3- Build Hadoop on Windows (one node)

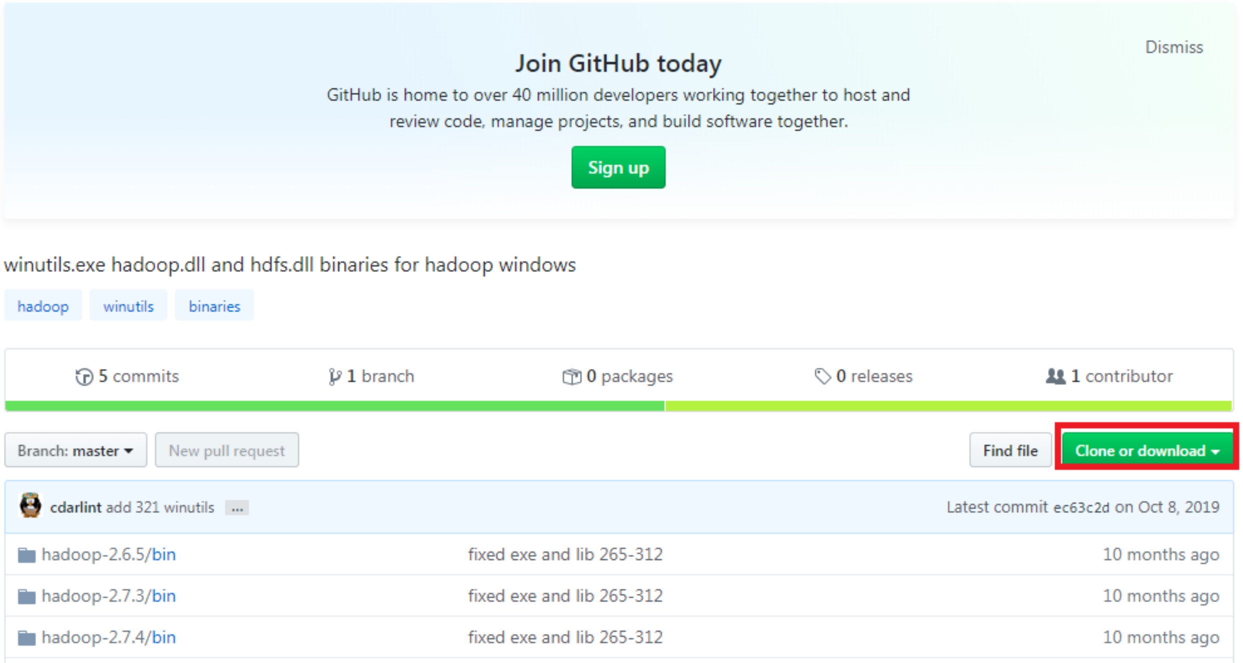
A- Download and install https://www.7-zip.org/ (to unzip Linux)

B- Download http://archive.apache.org/dist/hadoop/common/hadoop-3.1.2/hadoop-3.1.2.tar.gz

C- Make directory “bigdata” on c drive
C:\bigdata

D- Unzip hadoop-3.1.2.tar.gz
C:\bigdata\hadoop-3.1.2

E- Download Hadoop windows patch https://github.com/cdarlint/winutils



Join GitHub today

GitHub is home to over 40 million developers working together to host and review code, manage projects, and build software together.

winutils.exe hadoop.dll and hdfs.dll binaries for hadoop windows

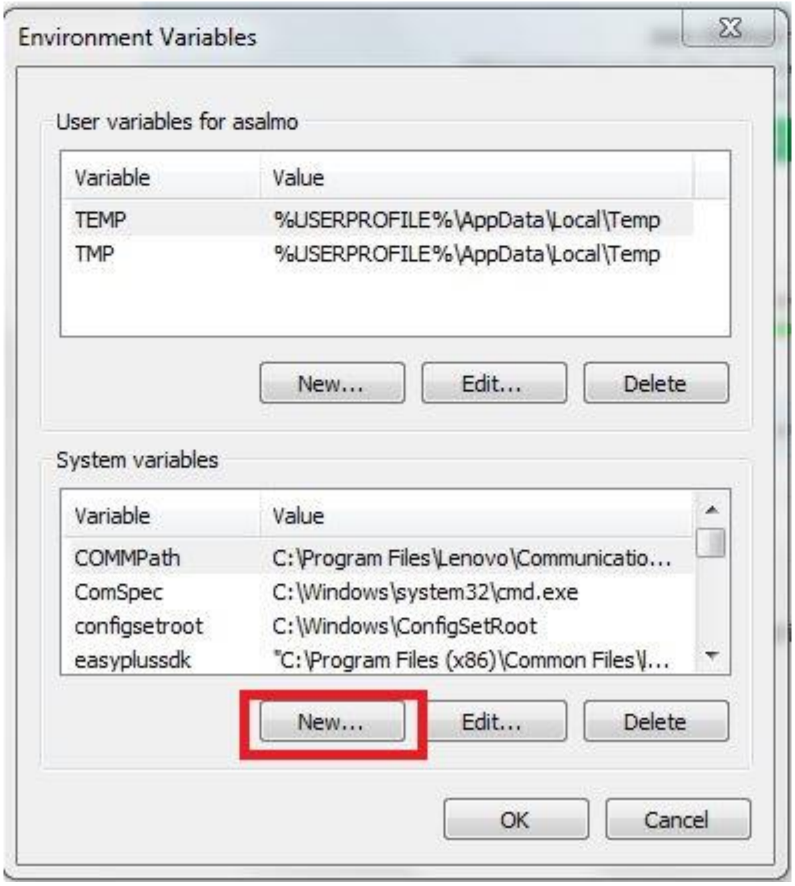
cdarlint add 321 winutils … Latest commit ec63c2d on Oct 8, 2019

hadoop-2.6.5/bin	fixed exe and lib 265-312	10 months ago
hadoop-2.7.3/bin	fixed exe and lib 265-312	10 months ago
hadoop-2.7.4/bin	fixed exe and lib 265-312	10 months ago

F- Download all patches, unzip the folder and copy Hadoop-3.1.2/bin to bin folder (C:\bigdata\hadoop-3.1.2\bin)

G- Build Hadoop variables:

Go “System properties” → Choose “Environment variables”



User variables for asalmo

Variable	Value
TEMP	%USERPROFILE%\AppData\Local\Temp
TMP	%USERPROFILE%\AppData\Local\Temp

System variables

Variable	Value
COMMPath	C:\Program Files\Lenovo\Communicatio...
ComSpec	C:\Windows\system32\cmd.exe
configsetroot	C:\Windows\ConfigSetRoot
easyplussdk	"C:\Program Files (x86)\Common Files\I...

H- Press “New” to add:

- HADOOP_HOME = C:\bigdata\ hadoop-3.1.2
- HADOOP_BIN = C:\bigdata\ hadoop-3.1.2\bin
- HADOOP_SBIN= C:\bigdata\ hadoop-3.1.2\sbin
- Add % HADOOP_HOME %; % HADOOP_BIN %;% HADOOP_SBIN% to **path** variable

I- Configure Hadoop to run on single machine
We will need to change the following files (C:\ bigdata\hadoop-2.9.1\etc\hadoop):
hadoop-env.cmd
core-site.xml
hdfs-site.xml
mapred-site.xml

1- hadoop-env.cmd:
Change JAVA_HOME variable

From:
set JAVA_HOME=%JAVA_HOME%
To:
set JAVA_HOME=C:\AdoptOpenJDK\jdk-8.0.232.09-hotspot

2- core-site.xml
Open this file to add:

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

Note: Replace the empty < configuration></ configuration>

3- hdfs-site.xml
Open this file to add:

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>file:///C:/bigdata/hadoop-3.1.2/data/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>file:///C:/bigdata/hadoop-3.1.2/data/datanode</value>
  </property>
</configuration>
```

Note: Replace the empty < configuration></ configuration>

In this case, we will need to make directory for

C:\BigData\ hadoop-3.1.2**data**
C:\BigData\ hadoop-3.1.2\data\namenode
C:\BigData\ hadoop-3.1.2\data\datanode

4- mapred-site.xml

Open this file to add:

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

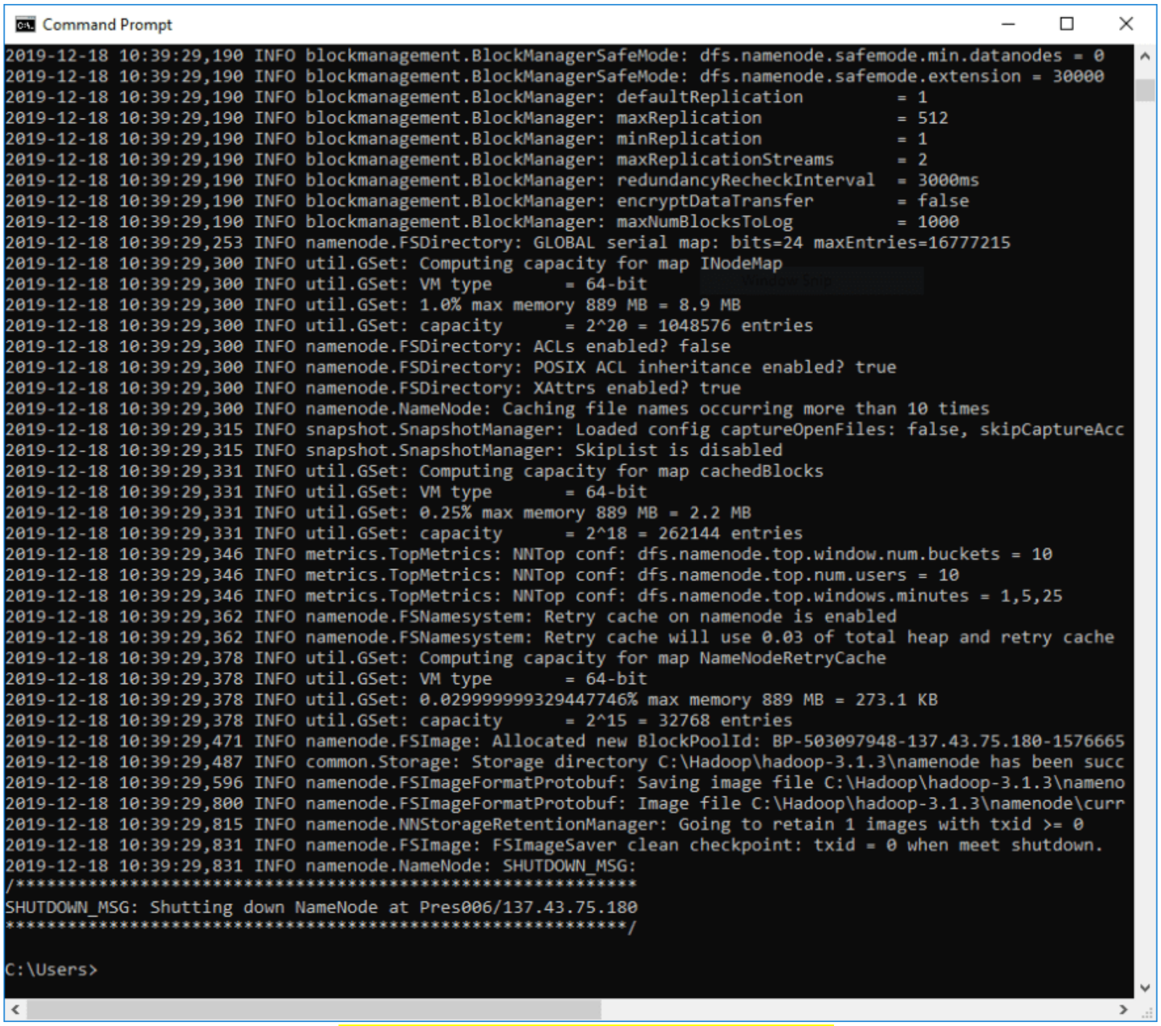
Note:

1- Replace the empty < configuration></ configuration>

J- Formant NameNode:

-Open cmd

-Type “hadoop namenode -format”



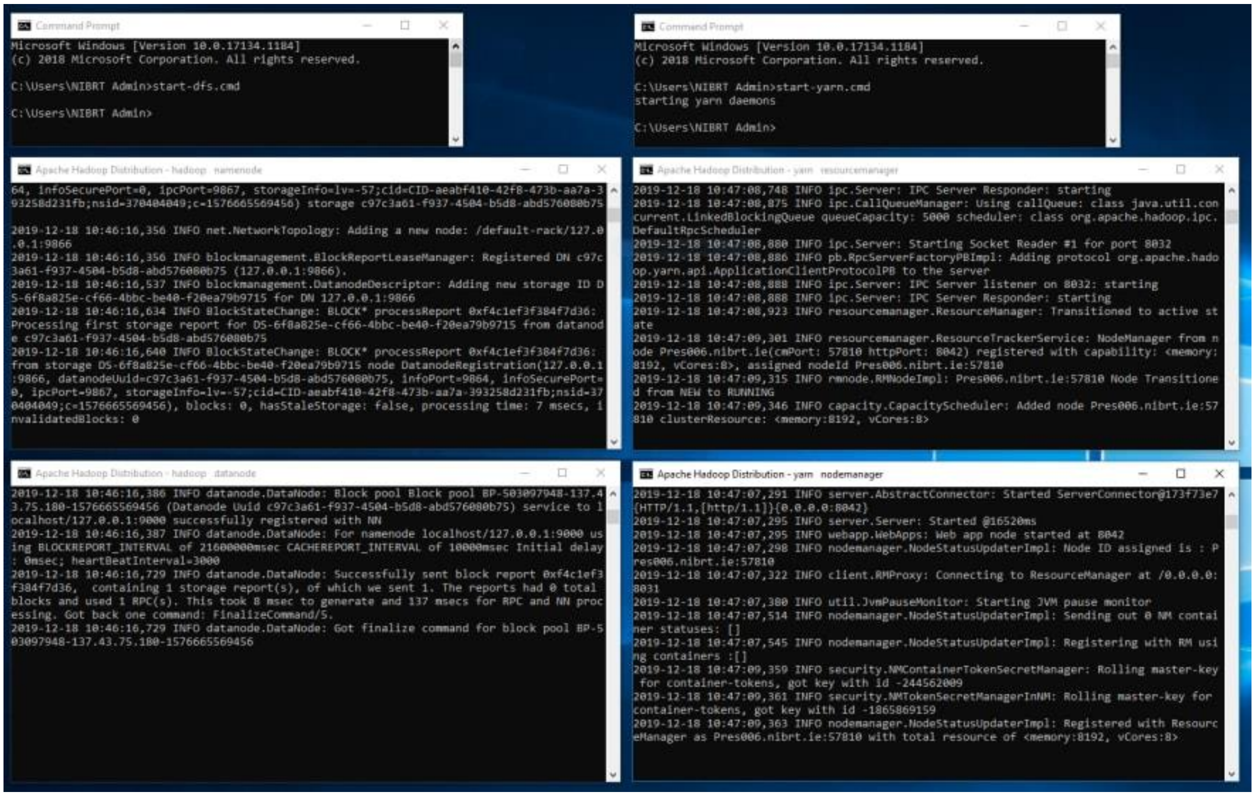
```
2019-12-18 10:39:29,190 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safemode.min.datanodes = 0
2019-12-18 10:39:29,190 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safemode.extension = 30000
2019-12-18 10:39:29,190 INFO blockmanagement.BlockManager: defaultReplication         = 1
2019-12-18 10:39:29,190 INFO blockmanagement.BlockManager: maxReplication             = 512
2019-12-18 10:39:29,190 INFO blockmanagement.BlockManager: minReplication             = 1
2019-12-18 10:39:29,190 INFO blockmanagement.BlockManager: maxReplicationStreams      = 2
2019-12-18 10:39:29,190 INFO blockmanagement.BlockManager: redundancyRecheckInterval  = 3000ms
2019-12-18 10:39:29,190 INFO blockmanagement.BlockManager: encryptDataTransfer        = false
2019-12-18 10:39:29,190 INFO blockmanagement.BlockManager: maxNumBlocksToLog          = 1000
2019-12-18 10:39:29,253 INFO namenode.FSDirectory: GLOBAL serial map: bits=24 maxEntries=16777215
2019-12-18 10:39:29,300 INFO util.GSet: Computing capacity for map INodeMap
2019-12-18 10:39:29,300 INFO util.GSet: VM type       = 64-bit
2019-12-18 10:39:29,300 INFO util.GSet: 1.0% max memory 889 MB = 8.9 MB
2019-12-18 10:39:29,300 INFO util.GSet: capacity      = 2^20 = 1048576 entries
2019-12-18 10:39:29,300 INFO namenode.FSDirectory: ACLs enabled? false
2019-12-18 10:39:29,300 INFO namenode.FSDirectory: POSIX ACL inheritance enabled? true
2019-12-18 10:39:29,300 INFO namenode.FSDirectory: XAttrs enabled? true
2019-12-18 10:39:29,300 INFO namenode.NameNode: Caching file names occurring more than 10 times
2019-12-18 10:39:29,315 INFO snapshot.SnapshotManager: Loaded config captureOpenFiles: false, skipCaptureAcc
2019-12-18 10:39:29,315 INFO snapshot.SnapshotManager: SkipList is disabled
2019-12-18 10:39:29,331 INFO util.GSet: Computing capacity for map cachedBlocks
2019-12-18 10:39:29,331 INFO util.GSet: VM type       = 64-bit
2019-12-18 10:39:29,331 INFO util.GSet: 0.25% max memory 889 MB = 2.2 MB
2019-12-18 10:39:29,331 INFO util.GSet: capacity      = 2^18 = 262144 entries
2019-12-18 10:39:29,346 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2019-12-18 10:39:29,346 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2019-12-18 10:39:29,346 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2019-12-18 10:39:29,362 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2019-12-18 10:39:29,362 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache
2019-12-18 10:39:29,378 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2019-12-18 10:39:29,378 INFO util.GSet: VM type       = 64-bit
2019-12-18 10:39:29,378 INFO util.GSet: 0.029999999329447746% max memory 889 MB = 273.1 KB
2019-12-18 10:39:29,378 INFO util.GSet: capacity      = 2^15 = 32768 entries
2019-12-18 10:39:29,471 INFO namenode.FSImage: Allocated new BlockPoolId: BP-503097948-137.43.75.180-1576665
2019-12-18 10:39:29,487 INFO common.Storage: Storage directory C:\Hadoop\hadoop-3.1.3\namenode has been succ
2019-12-18 10:39:29,596 INFO namenode.FSImageFormatProtobuf: Saving image file C:\Hadoop\hadoop-3.1.3\nameno
2019-12-18 10:39:29,800 INFO namenode.FSImageFormatProtobuf: Image file C:\Hadoop\hadoop-3.1.3\namenode\curr
2019-12-18 10:39:29,815 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2019-12-18 10:39:29,831 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid = 0 when meet shutdown.
2019-12-18 10:39:29,831 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at Pres006/137.43.75.180
************************************************************/

C:\Users>
```

K- One more thing to do: copy hadoop-yarn-server-timelineservice-3.1.2
from C:\bigdata\hadoop-3.1.2\share\hadoop\yarn\timelineservice
to C:\bigdata\hadoop-3.1.2\share\hadoop\yarn

L- We need to type start-all.cmd to start all nodes on one machine

You should have:



Command Prompt
Command Prompt
Apache Hadoop Distribution - hadoop namenode
Apache Hadoop Distribution - yarn resourcemanager
Apache Hadoop Distribution - hadoop datanode
Apache Hadoop Distribution - yarn nodemanager

```
Microsoft Windows [Version 10.0.17134.1184]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\NIBRT Admin>start-dfs.cmd

C:\Users\NIBRT Admin>
```

```
Microsoft Windows [Version 10.0.17134.1184]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\NIBRT Admin>start-yarn.cmd
starting yarn daemons

C:\Users\NIBRT Admin>
```

```
64, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-aeabf410-42f8-473b-aa7a-3
93258d231fb;nsid=370404049;c=1576665569456) storage c97c3a61-f937-4504-b5d8-abd576080b75

2019-12-18 10:46:16,356 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0
.0.1:9866
2019-12-18 10:46:16,356 INFO blockmanagement.BlockReportLeaseManager: Registered DN c97c
3a61-f937-4504-b5d8-abd576080b75 (127.0.0.1:9866).
2019-12-18 10:46:16,537 INFO blockmanagement.DatanodeDescriptor: Adding new storage ID D
S-6f8a825e-cf66-4bbc-be40-f20ea79b9715 for DN 127.0.0.1:9866
2019-12-18 10:46:16,634 INFO BlockStateChange: BLOCK* processReport 0xf4c1ef3f384f7d36:
Processing first storage report for DS-6f8a825e-cf66-4bbc-be40-f20ea79b9715 from datanod
e c97c3a61-f937-4504-b5d8-abd576080b75
2019-12-18 10:46:16,640 INFO BlockStateChange: BLOCK* processReport 0xf4c1ef3f384f7d36:
from storage DS-6f8a825e-cf66-4bbc-be40-f20ea79b9715 node DatanodeRegistration(127.0.0.1
:9866, datanodeUuid=c97c3a61-f937-4504-b5d8-abd576080b75, infoPort=9864, infoSecurePort=
0, ipcPort=9867, storageInfo=lv=-57;cid=CID-aeabf410-42f8-473b-aa7a-393258d231fb;nsid=37
0404049;c=1576665569456), blocks: 0, hasStaleStorage: false, processing time: 7 msecs, i
nvalidatedBlocks: 0
```

```
2019-12-18 10:47:08,748 INFO ipc.Server: IPC Server Responder: starting
2019-12-18 10:47:08,875 INFO ipc.CallQueueManager: Using callQueue: class java.util.con
current.LinkedBlockingQueue queueCapacity: 5000 scheduler: class org.apache.hadoop.ipc.
DefaultRpcScheduler
2019-12-18 10:47:08,880 INFO ipc.Server: Starting Socket Reader #1 for port 8032
2019-12-18 10:47:08,886 INFO pb.RpcServerFactoryPBImpl: Adding protocol org.apache.hado
op.yarn.api.ApplicationClientProtocolPB to the server
2019-12-18 10:47:08,888 INFO ipc.Server: IPC Server listener on 8032: starting
2019-12-18 10:47:08,888 INFO ipc.Server: IPC Server Responder: starting
2019-12-18 10:47:08,923 INFO resourcemanager.ResourceManager: Transitioned to active st
ate
2019-12-18 10:47:09,301 INFO resourcemanager.ResourceTrackerService: NodeManager from n
ode Pres006.nibrt.ie(cmPort: 57810 httpPort: 8042) registered with capability: <memory:
8192, vCores:8>, assigned nodeId Pres006.nibrt.ie:57810
2019-12-18 10:47:09,315 INFO rmnode.RMNodeImpl: Pres006.nibrt.ie:57810 Node Transitione
d from NEW to RUNNING
2019-12-18 10:47:09,346 INFO capacity.CapacityScheduler: Added node Pres006.nibrt.ie:57
810 clusterResource: <memory:8192, vCores:8>
```

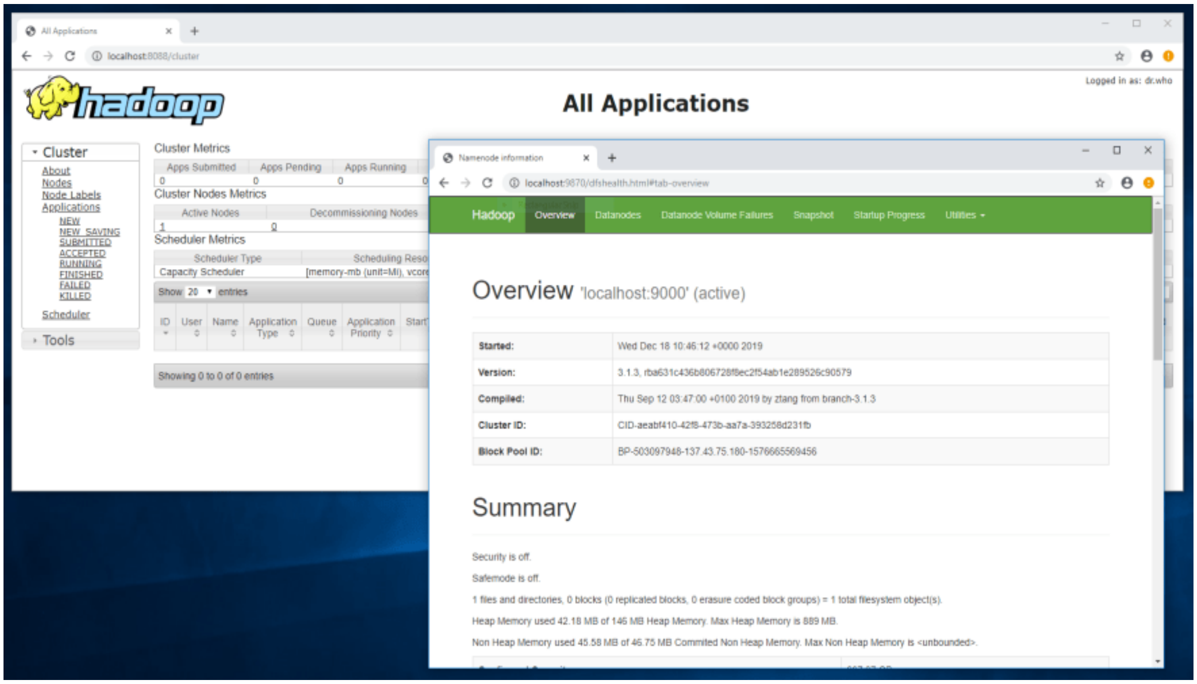
```
2019-12-18 10:46:16,386 INFO datanode.DataNode: Block pool Block pool BP-503097948-137.4
3.75.180-1576665569456 (Datanode Uuid c97c3a61-f937-4504-b5d8-abd576080b75) service to l
ocalhost/127.0.0.1:9000 successfully registered with NN
2019-12-18 10:46:16,387 INFO datanode.DataNode: For namenode localhost/127.0.0.1:9000 us
ing BLOCKREPORT_INTERVAL of 21600000msec CACHEREPORT_INTERVAL of 10000msec Initial delay
: 0msec; heartBeatInterval=3000
2019-12-18 10:46:16,729 INFO datanode.DataNode: Successfully sent block report 0xf4c1ef3
f384f7d36,  containing 1 storage report(s), of which we sent 1. The reports had 0 total
blocks and used 1 RPC(s). This took 8 msec to generate and 137 msecs for RPC and NN proc
essing. Got back one command: FinalizeCommand/5.
2019-12-18 10:46:16,729 INFO datanode.DataNode: Got finalize command for block pool BP-5
03097948-137.43.75.180-1576665569456
```

```
2019-12-18 10:47:07,291 INFO server.AbstractConnector: Started ServerConnector@173f73e7
{HTTP/1.1,[http/1.1]}{0.0.0.0:8042}
2019-12-18 10:47:07,295 INFO server.Server: Started @16520ms
2019-12-18 10:47:07,295 INFO webapp.WebApps: Web app node started at 8042
2019-12-18 10:47:07,298 INFO nodemanager.NodeStatusUpdaterImpl: Node ID assigned is : P
res006.nibrt.ie:57810
2019-12-18 10:47:07,322 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:
8031
2019-12-18 10:47:07,380 INFO util.JvmPauseMonitor: Starting JVM pause monitor
2019-12-18 10:47:07,514 INFO nodemanager.NodeStatusUpdaterImpl: Sending out 0 NM contai
ner statuses: []
2019-12-18 10:47:07,545 INFO nodemanager.NodeStatusUpdaterImpl: Registering with RM usi
ng containers :[]
2019-12-18 10:47:09,359 INFO security.NMContainerTokenSecretManager: Rolling master-key
 for container-tokens, got key with id -244562009
2019-12-18 10:47:09,361 INFO security.NMTokenSecretManagerInNM: Rolling master-key for
container-tokens, got key with id -1865869159
2019-12-18 10:47:09,363 INFO nodemanager.NodeStatusUpdaterImpl: Registered with Resourc
eManager as Pres006.nibrt.ie:57810 with total resource of <memory:8192, vCores:8>
```

Now you should have Hadoop on your machine

M- Hadoop Web UI

You can also open http://localhost:8088 and http://localhost:9870 in your browser



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running
0	0	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes
1	0

Scheduler Metrics

Scheduler Type	Scheduling Reso
Capacity Scheduler	[memory-mb (unit=Mi), vcore

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	Start

Showing 0 to 0 of 0 entries

Overview 'localhost:9000' (active)

Started:	Wed Dec 18 10:46:12 +0000 2019
Version:	3.1.3, rba631c436b806728f8ec2f54ab1e289526c90579
Compiled:	Thu Sep 12 03:47:00 +0100 2019 by ztang from branch-3.1.3
Cluster ID:	CID-aeabf410-42f8-473b-aa7a-393258d231fb
Block Pool ID:	BP-503097948-137.43.75.180-1576665569456

Summary

Security is off.

Safemode is off.

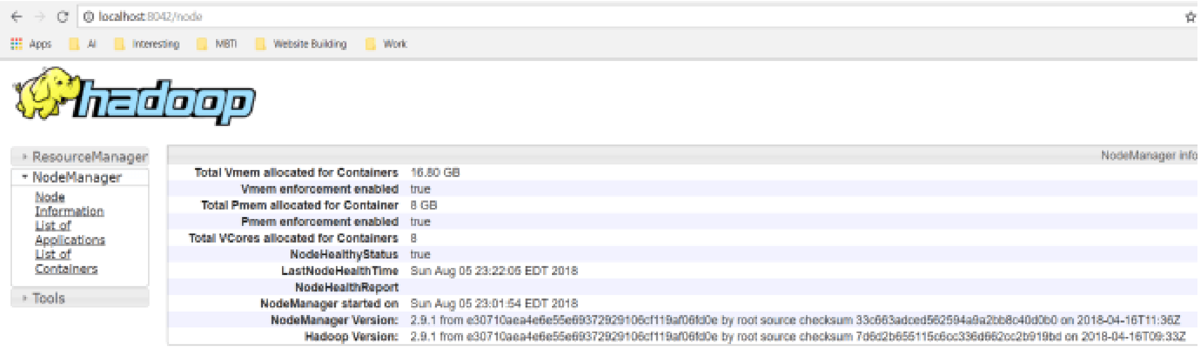
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Heap Memory used 42.18 MB of 146 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 45.58 MB of 46.75 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

1- Node Manager

http://localhost:8042



	NodeManager information
Total Vmem allocated for Containers	16.80 GB
Vmem enforcement enabled	true
Total Pmem allocated for Container	8 GB
Pmem enforcement enabled	true
Total VCores allocated for Containers	8
NodeHealthyStatus	true
LastNodeHealthTime	Sun Aug 05 23:22:06 EDT 2018
NodeHealthReport	
NodeManager started on	Sun Aug 05 23:01:54 EDT 2018
NodeManager Version:	2.9.1 from e30710aea4e6e55e69372929106cf119af06fd0e by root source checksum 33c663adced562594a9a2bb8c40d0b0 on 2018-04-16T11:36Z
Hadoop Version:	2.9.1 from e30710aea4e6e55e69372929106cf119af06fd0e by root source checksum 7d6d2b655115c6cc336d662cc2b919bd on 2018-04-16T09:33Z

K- Working with HDFS

```
>notepad Sample.txt
```

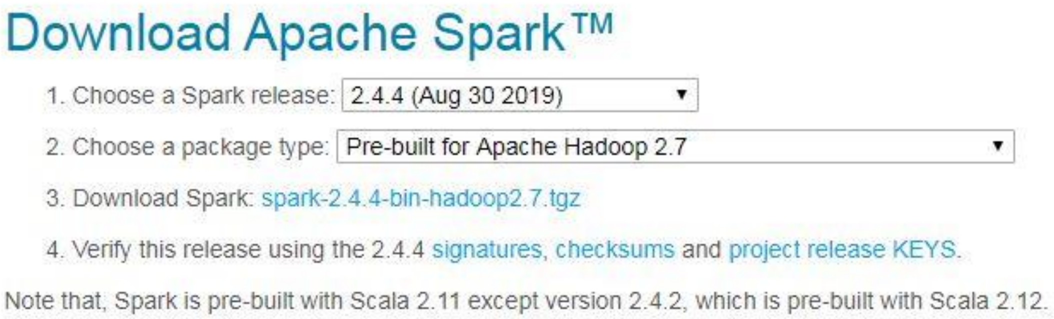
Write anything and save the file

```
hdfs dfs -ls /
hdfs dfs -mkdir /test
hdfs dfs -copyFromLocal Sample.txt /test
hdfs dfs -ls /test
hdfs dfs -cat /test/Sample.txt
```

1- Hadoop reference: https://dev.to/awwsmm/installing-and-running-hadoop-and-spark-on-windows-33kc#comments
2- Hadoop Patch reference https://github.com/cdarlint/winutils

Install Spark

A. Download Spark version 2.3 or 2.4 (https://spark.apache.org/downloads.html)



Download Apache Spark™

1. Choose a Spark release: 2.4.4 (Aug 30 2019) ▾
2. Choose a package type: Pre-built for Apache Hadoop 2.7 ▾
3. Download Spark: spark-2.4.4-bin-hadoop2.7.tgz
4. Verify this release using the 2.4.4 signatures, checksums and project release KEYS.

Note that, Spark is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12.

B. Unzip spark-2.4.4-bin-hadoop2.7.tgz

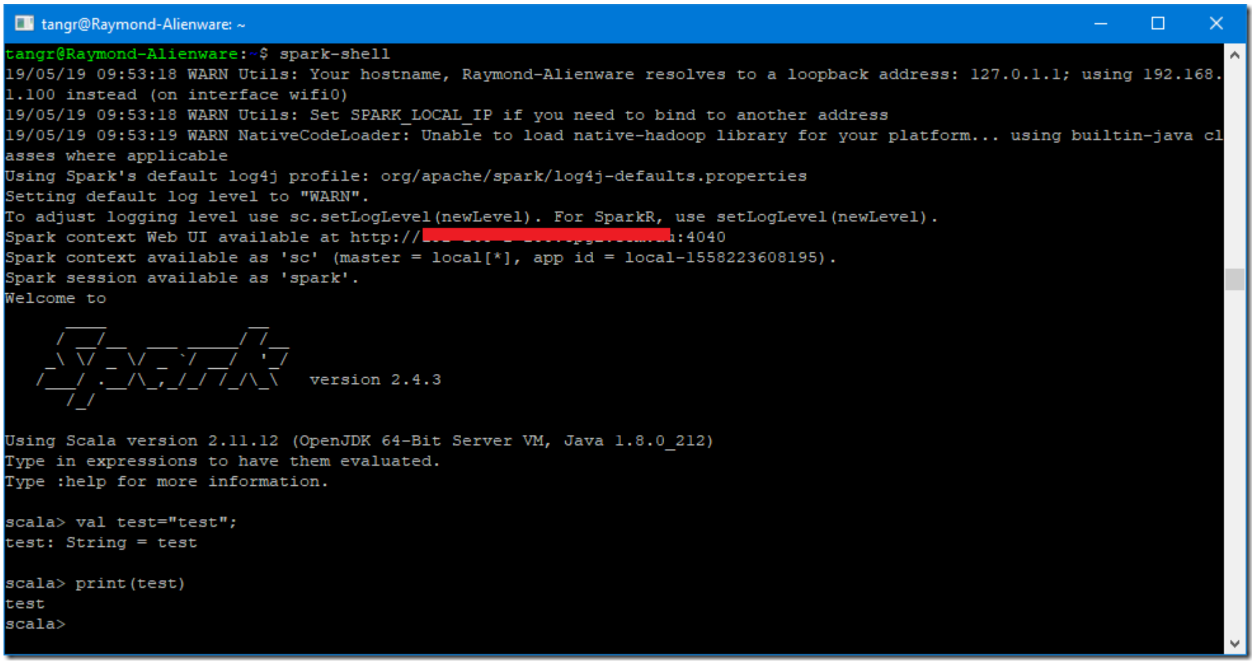
C. Put the folder in C:\bigdata\spark-2.4.4-bin-hadoop2.7

D. Add to system variables
SPARK_HOME=C:\alaa\bigdata\spark-2.4.4-bin-hadoop2.7
PYSPARK_PYTHON=C:\Users\asalmo\AppData\Local\Programs\Python\Python37\python.exe
PYSPARK_DRIVER_PYTHON=
C:\Users\asalmo\AppData\Local\Programs\Python\Python37\python.exe

E. Add SPARK_HOME to path % SPARK_HOME %\bin

F. Open cmd

G. Type: spark-shell to start spark scala



```
tangr@Raymond-Alienware:~$ spark-shell
19/05/19 09:53:18 WARN Utils: Your hostname, Raymond-Alienware resolves to a loopback address: 127.0.1.1; using 192.168.
1.100 instead (on interface wifi0)
19/05/19 09:53:18 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
19/05/19 09:53:19 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
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).
Spark context Web UI available at http://[redacted]:4040
Spark context available as 'sc' (master = local[*], app id = local-1558223608195).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.3
      /_/

Using Scala version 2.11.12 (OpenJDK 64-Bit Server VM, Java 1.8.0_212)
Type in expressions to have them evaluated.
Type :help for more information.

scala> val test="test";
test: String = test

scala> print(test)
test
scala>
```

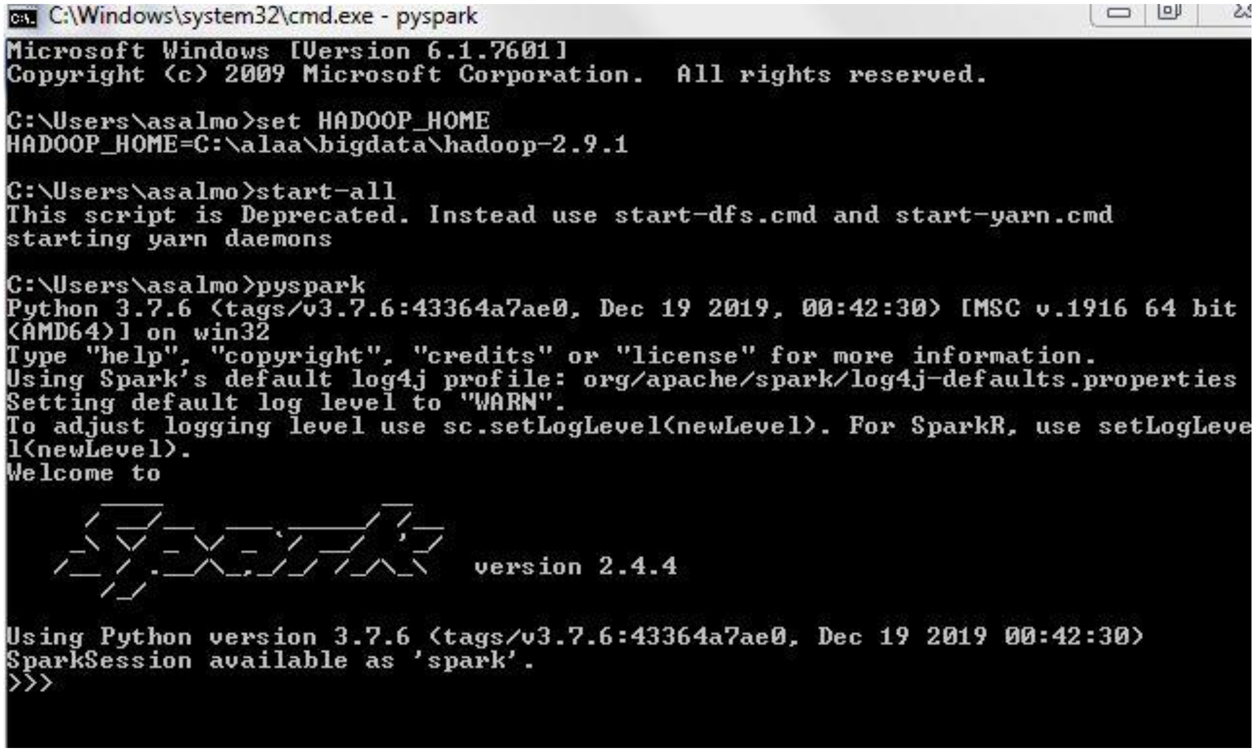
H. Exit type :q

I. You need to install Python (download: https://www.python.org/downloads/windows/)

J. Add the variable: PYSPARK_DRIVER_PYTHON

PYSPARK_DRIVER_PYTHON= C:\Users\XXXXXXXX (Usename)\AppData\Local\Programs\Python\Python37\python.exe

K. Open cmd and type pyspark



```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Users\asalmo>set HADOOP_HOME
HADOOP_HOME=C:\alaa\bigdata\hadoop-2.9.1

C:\Users\asalmo>start-all
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\Users\asalmo>pyspark
Python 3.7.6 (tags/v3.7.6:43364a7ae0, Dec 19 2019, 00:42:30) [MSC v.1916 64 bit
(AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
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 setLogLeve
l(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.4
      /_/

Using Python version 3.7.6 (tags/v3.7.6:43364a7ae0, Dec 19 2019 00:42:30)
SparkSession available as 'spark'.
>>>
```

N- To quit write quit()

Reference: https://kontext.tech/column/spark/311/apache-spark-243-installation-on-windows-10-using-windows-subsystem-for-linux

Anaconda with Jupyter notebook

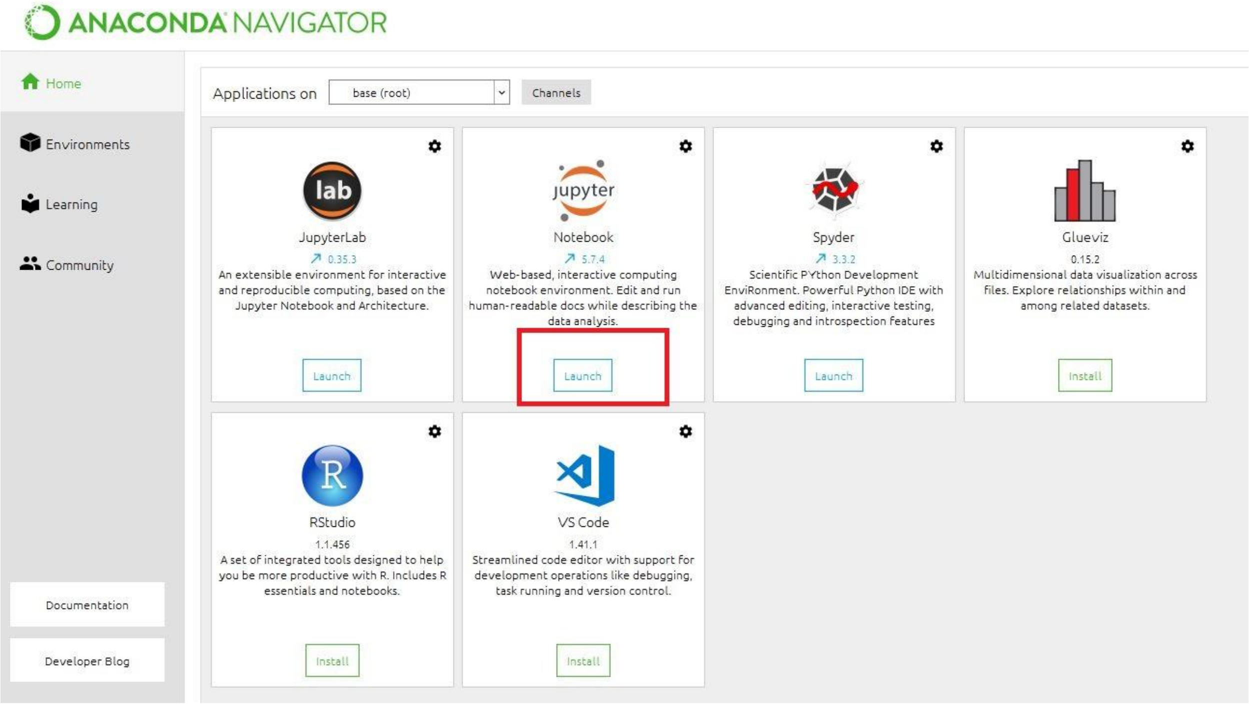
1- Install Anaconda

Note: For Windows 7 install Anaconda with 2018 and Python 3.7
For Windows 10 install Anaconda with 2019 and Python 3.7

https://docs.anaconda.com/anaconda/packages/oldpkglists/

2- After installation,

Go to window search for anaconda



Applications on base (root)

JupyterLab
0.35.3
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Notebook
5.7.4
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Spyder
3.3.2
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Glueviz
0.15.2
Multidimensional data visualization across files. Explore relationships within and among related datasets.

RStudio
1.1.456
A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.

VS Code
1.41.1
Streamlined code editor with support for development operations like debugging, task running and version control.

Press Jupyter to start the web notebook.

3- Install findspark

Go to “search program and files” write “anaconda”

Choose “Anaconda Prompt”

```
Write "conda install -c conda-forge findspark"
```

```
4- Install Pandas
```

Go to “search program and files” write “anaconda”
Choose “Anaconda Prompt”

```
Wite "pip install pandas"
```