INSTITUTE OF TECHNOLOGY

SCHOOL OF COMPUTING

DEPARTMENT OF SOFTWARE ENGINEERING

POSTGRADUATE PROGRAM

Big Data Analytics for Water Consumption

By:

1. **Ashagrew Liyih**
2. **Demeke Aschale**
3. **Derso Dessie**

Instructor: Zewdie Mossie (Ph.D.**)**

April, 2023
Debre Markos, Ethiopia

Software Requirement

- Download Installing Spark -3.0.1-bin-hadoop2.7.
- Download winutils exe files
- Download and installing jdk1.8.0_111_windows-x64_bin.
- Installing Anaconda 3.7
- Installing python 3.8 and add path to the enviroment

Setting Environmental variable

We have to set environment variables for handoop ,java ,spark. First, copy the winutils exe file to the apache spark bin folder as shown below.

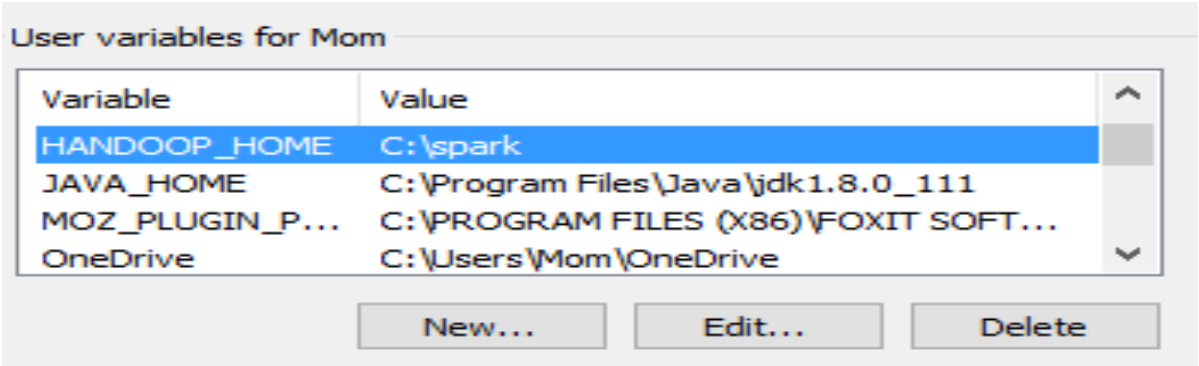
spark-shell	8/28/2020 5:10 AM	File	4 KB
spark-shell	8/28/2020 5:10 AM	Windows Comma...	2 KB
spark-shell2	8/28/2020 5:10 AM	Windows Comma...	2 KB
spark-sql	8/28/2020 5:10 AM	File	2 KB
spark-sql	8/28/2020 5:10 AM	Windows Comma...	2 KB
spark-sql2	8/28/2020 5:10 AM	Windows Comma...	2 KB
spark-submit	8/28/2020 5:10 AM	File	2 KB
spark-submit	8/28/2020 5:10 AM	Windows Comma...	2 KB
spark-submit2	8/28/2020 5:10 AM	Windows Comma...	2 KB
winutils	12/26/2020 8:57 AM	Application	107 KB

Second ,copy apache spark folder to C:// directory as shown below

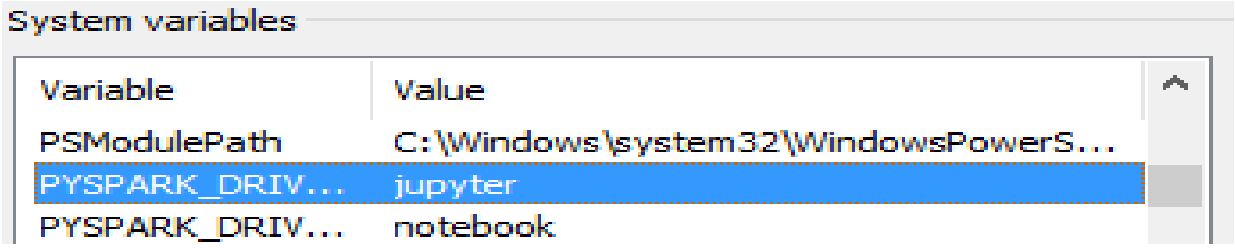
This PC > Local Disk (C:)

Name	Date modified	Type	Size
Autodesk	10/26/2022 7:27 AM	File folder	
DRIVERS	9/28/2022 9:41 AM	File folder	
Intel	9/23/2022 7:43 AM	File folder	
PerfLogs	7/10/2015 8:04 AM	File folder	
Program Files	3/21/2023 10:23 AM	File folder	
Program Files (x86)	4/19/2023 2:09 AM	File folder	
SPARK	3/25/2023 2:37 AM	File folder	
tmp	4/4/2023 4:31 AM	File folder	
Users	11/28/2022 6:28 PM	File folder	
usr	12/12/2022 6:57 AM	File folder	
Windows	3/22/2023 5:43 AM	File folder	
xampp	9/23/2022 10:11 AM	File folder	

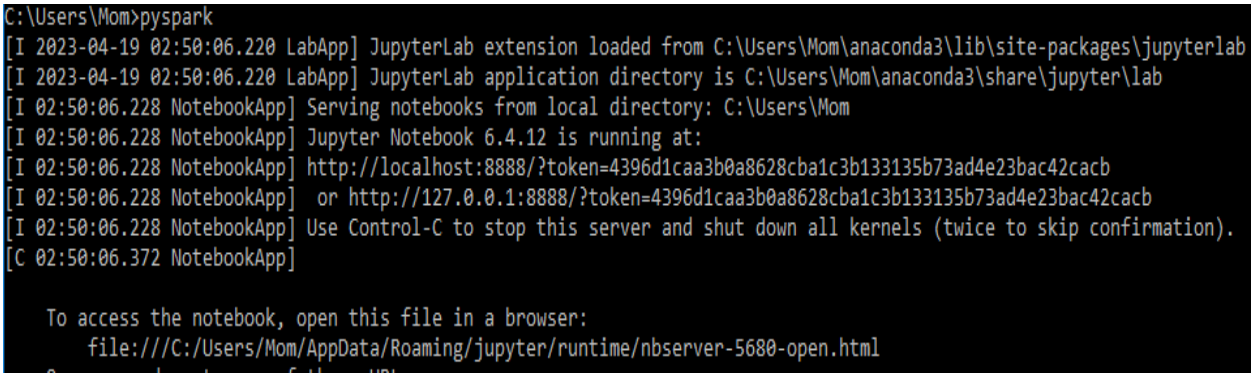
Finally, set the environment variables for the spark and handoop as shown below.



Install anaconda and integrate apache spark to the jupyter notebook with the environment variable below.



After installing and set the environment open command prompt ,and type “pyspark” and check whether the Apache spark integrate with jupyter or not.



```
C:\Users\Mom>pyspark
[I 2023-04-19 02:50:06.220 LabApp] JupyterLab extension loaded from C:\Users\Mom\anaconda3\lib\site-packages\jupyterlab
[I 2023-04-19 02:50:06.220 LabApp] JupyterLab application directory is C:\Users\Mom\anaconda3\share\jupyter\lab
[I 02:50:06.228 NotebookApp] Serving notebooks from local directory: C:\Users\Mom
[I 02:50:06.228 NotebookApp] Jupyter Notebook 6.4.12 is running at:
[I 02:50:06.228 NotebookApp] http://localhost:8888/?token=4396d1caa3b0a8628cba1c3b133135b73ad4e23bac42cacb
[I 02:50:06.228 NotebookApp]  or http://127.0.0.1:8888/?token=4396d1caa3b0a8628cba1c3b133135b73ad4e23bac42cacb
[I 02:50:06.228 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 02:50:06.372 NotebookApp]

    To access the notebook, open this file in a browser:
        file:///C:/Users/Mom/AppData/Roaming/jupyter/runtime/nbserver-5680-open.html
```

Data preparation

- Download the dataset in the link below.
  https://www.kaggle.com/datasets/marcomolina/water-consumption-in-a-median-size-city?resource=download&select=AguaH.csv.
- Read the csv file and create correlation between each column to check the columns are important for analytics or not.

```
## read the csv file
df=spark.read.csv(
                path="AguaH.csv",
                  header=True,
                  inferSchema=True,
                )
```

UL	1	0.28	0.3	0.3	0.3	0.28	0.28	0.28	0.26	0.27	0.26	0.26
ENE_09	0.28	1	0.98	0.96	0.97	0.95	0.99	0.99	0.99	0.99	0.99	0.99
FEB_09	0.3	0.98	1	0.94	0.95	0.9	0.95	0.95	0.96	0.96	0.96	0.95
MAR_09	0.3	0.96	0.94	1	1	0.99	0.98	0.97	0.95	0.95	0.93	0.96
ABR_09	0.3	0.97	0.95	1	1	0.99	0.98	0.98	0.97	0.96	0.95	0.97
MAY_09	0.28	0.95	0.9	0.99	0.99	1	0.98	0.98	0.96	0.95	0.94	0.97
JUN_09	0.28	0.99	0.95	0.98	0.98	0.98	1	1	0.99	0.99	0.98	1
JUL_09	0.28	0.99	0.95	0.97	0.98	0.98	1	1	1	0.99	0.99	1
AGO_09	0.26	0.99	0.96	0.95	0.97	0.96	0.99	1	1	1	1	1

All of the columns are positively correlated .Therefore, we are not able to remove the column and simply fill 0 if it contains null values as shown below.

```
### fill the null values as 0 and count total number of records in the dataset
dfill=df701.na.fill(0)
dcount=dfill.count()
print(dcount)
```

```
178597
```

Task to be implemented

- Load the csv file and count number of records
- Create correlation of the column
- Fill the record if it is null
- Find mean ,max and min of each column with TU=**COMERCIAL**,**SOCIAL**,**INDUSTRIAL** and count records with the three TU values using spark and pandas
- Find standard deviations and variances of each column
- Min Max Normalization using min-max normalization in a range of 0 and 1
- Visualize the dataset