# M2\_data\_analysis\_platform

INSTALLATION GUIDE

Installation steps in Windows:

1. Clone repository <https://github.com/Metropolis-2/M2_data_analysis_platform.git> and checkout to branch [**develop\_everis**] or last stable branch by default.
2. Download a suitable IDE like **Pycharm** (<https://www.jetbrains.com/es-es/pycharm/> ) or **Spyder** (<https://www.spyder-ide.org/> ).

It is recommended to download the **Jetbrains Toolbox** (<https://www.jetbrains.com/es-es/toolbox-app/> ) that allows you to manage different IDEs easily. You can also download **Anaconda** (<https://www.anaconda.com/> ) to manage these tools through an application dashboard. It also allows you to manage conda environments visually.

If a more lightweight manager is needed, you can download **Miniconda** (<https://docs.conda.io/en/latest/miniconda.html> ) and manage your conda environments directly by console.

1. Load the **M2\_data\_analysis\_platform** project into IDE.
2. Install **Java 8** (<https://www.oracle.com/java/technologies/downloads/#java8-windows>)
3. Post installation, set **JAVA\_HOME** and **PATH** variable.

* *JAVA\_HOME = C:\Program Files\Java\jdk1.8.0\_321*
* *PATH = %PATH%;C:\Program Files\Java\jdk1.8.0\_321\bin*

1. Download **Apache Spark** (<https://spark.apache.org/downloads.html> ). Choose last Spark release and corresponding package type. We recommend: “*Spark 3.2.1 (Jan 26 2022)*” and package “*Pre-build for Apache Hadoop 3.3 and later*”. Download [spark-3.2.1-bin-hadoop3.2.tgz](https://www.apache.org/dyn/closer.lua/spark/spark-3.2.1/spark-3.2.1-bin-hadoop3.2.tgz) and unzip **spark-3.2.1-bin-hadoop3.2** folder into user folder (for example: *C:\Users\username\apps*) Now set the following environment variables.

* *SPARK\_HOME = C:\apps\spark-3.2.1-bin-hadoop3.2*
* *HADOOP\_HOME = C:\apps\ spark-3.2.1-bin-hadoop3.2*
* *PATH=%PATH%;C:\apps\spark-3.2.1-bin-hadoop3.2\bin*
* *PYSPARK\_PYTHON = python*

1. Download **Apache Hadoop**, the framework for Spark, from <https://github.com/cdarlint/winutils>.

It is necessary to use the version corresponding to the type of Apache Spark package downloaded, in our case, in our case **hadoop-3.2** (<https://github.com/cdarlint/winutils/tree/master/hadoop-3.2.0/bin>)

once the bin folder is downloaded copy all the files in the spark bin folder **spark-3.2.1-bin-hadoop3.2** previously prepared.

1. **Create virtual environment with conda**. In the ***windows-spec-file.txt*** of the repository you can find the command to create an environment: ***conda create --name <env> --file <this file>***

Example: *conda create --name m2\_pyspark --file windows-spec-file.txt*

1. Once the virtual environment is created, import it from the IDE. For example, in PyCharm it is from the option: *Add Interpreter -> Conda Environment -> Existing environment -> Interpreter -> add path of your env*

Finally **activate the environment** and check that there are no apparent errors or problems.

Installation steps in Ubuntu:

1. Clone repository <https://github.com/Metropolis-2/M2_data_analysis_platform.git> and checkout to branch [**develop\_everis**] or last stable branch by default.
2. Download a suitable IDE like **Pycharm** (<https://www.jetbrains.com/es-es/pycharm/> ) or **Spyder** (<https://www.spyder-ide.org/> ).

It is recommended to download the **Jetbrains Toolbox** (<https://www.jetbrains.com/es-es/toolbox-app/> ) that allows you to manage different IDEs easily. You can also download **Anaconda** (<https://www.anaconda.com/> ) to manage these tools through an application dashboard. It also allows you to manage conda environments visually.

If a more lightweight manager is needed, you can download **Miniconda** (<https://docs.conda.io/en/latest/miniconda.html> ) and manage your conda environments directly by console.

1. Load the **M2\_data\_analysis\_platform** project into IDE.
2. Install **Java 8** with:
   * sudo apt install openjdk-8-jdk
3. Download **Apache Spark** (<https://spark.apache.org/downloads.html> ). Choose last Spark release and corresponding package type. We recommend: “*Spark 3.2.1 (Jan 26 2022)*” and package “*Pre-build for Apache Hadoop 3.3 and later*”. Download [spark-3.2.1-bin-hadoop3.2.tgz](https://www.apache.org/dyn/closer.lua/spark/spark-3.2.1/spark-3.2.1-bin-hadoop3.2.tgz) and unzip **spark-3.2.1-bin-hadoop3.2** folder into user folder (for example: /home/adrigrillo/.local/bin/spark-3.2.1-bin-hadoop3.2)
4. Once downloaded, it is needed to set the environment variables for Java and Spark. Therefore, opening the file .bashrc of your user (for example, ‘/home/adrigrillo/.bashrc’) append the following lines, modifying them with your paths:

* *export SPARK\_HOME=*/home/adrigrillo/.local/bin/spark-3.2.1-bin-hadoop3.2
* export PATH=$PATH:$SPARK\_HOME/bin
* export PYTHONPATH=$SPARK\_HOME/python:$PYTHONPATH
* export PYSPARK\_DRIVER\_PYTHON="jupyter"
* export PYSPARK\_DRIVER\_PYTHON\_OPTS="notebook"
* export PYSPARK\_PYTHON=python
* export HADOOP\_HOME=/home/adrigrillo/.local/bin/spark-3.2.1-bin-hadoop3.2
* export JAVA\_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
* export PATH=$PATH:$JAVA\_HOME/bin

1. **Create virtual environment with conda**. In the ***ubuntu-spec-file.txt*** of the repository you can find the command to create an environment: ***conda create --name <env> --file <this file>***

Example: *conda create --name m2\_pyspark --file ubuntu-spec-file.txt*

1. Once the virtual environment is created, import it from the IDE. For example, in PyCharm it is from the option: *Add Interpreter -> Conda Environment -> Existing environment -> Interpreter -> add path of your env*

Finally **activate the environment** and check that there are no apparent errors or problems.