Setup Apache Spark

Start an ubuntu machine in hyper V.

Update ubuntu machine.

```
Sudo apt update
Sudo apt upgrade
```

To get start with PySpark you need to do Manual Installation:

Install Java (Spark requires Java):

```
sudo apt install openjdk-11-jdk
```

Download Apache Spark:
Go to the Apache Spark downloads page and choose a version (e.g., Spark 3.5.0).
Copy the download link for the pre-built package for Hadoop.
Use wget to download it:
Extract the downloaded file:

```
tar -xzf spark-3.5.0-bin-hadoop3.tgz
```

Move it to a more appropriate directory (optional):

```
sudo mv spark-3.5.0-bin-hadoop3 /opt/spark
```

Set Environment Variables: Add the following lines to your ~/.bashrc or ~/.profile file:

```
export SPARK_HOME=/opt/spark
export PATH=$PATH:$SPARK_HOME/bin
```

Load the environment variables:

```
source ~/.bashrc
```

Verify the Installation: You can verify the installation by running:

```
spark-shell
```

Run the code with spark-submit yourcode.py

Here’s a typical command structure for using spark-submit:

```
spark-submit [options] your_script.py [script arguments]
```

Common Options

1. **--master**: Specifies the cluster manager to connect to (e.g., local, yarn, mesos).
    - Example: --master local[4] runs the job locally with 4 threads.
2. **--deploy-mode**: Indicates how to deploy the driver program (e.g., client or cluster).
    - Example: --deploy-mode cluster runs the driver on the cluster.
3. **--executor-memory**: Specifies the amount of memory to use per executor process.
    - Example: --executor-memory 2G allocates 2 GB of memory for each executor.
4. **--num-executors**: Defines the number of executors to launch.
    - Example: --num-executors 10 starts 10 executor instances.
5. **--py-files**: Distributes .zip or .py files to the worker nodes.
    - Example: --py-files my_package.zip includes additional Python packages.