Installing and Running Apache Spark on Ubuntu (From Scratch)

# Step 1: Install Java (OpenJDK 11)

Apache Spark requires Java to run. Use the following commands:

sudo apt update  
sudo apt install openjdk-11-jdk  
java -version

# Step 2: Download and Extract Apache Spark

Download Spark (pre-built for Hadoop 3):

wget https://dlcdn.apache.org/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz

Extract the file and move it to /opt directory:

tar -xvzf spark-3.5.0-bin-hadoop3.tgz  
sudo mv spark-3.5.0-bin-hadoop3 /opt/spark

# Step 3: Set Environment Variables

Add the following lines to your ~/.bashrc file to set the necessary environment variables:

nano ~/.bashrc

export SPARK\_HOME=/opt/spark  
export PATH=$PATH:$SPARK\_HOME/bin  
export JAVA\_HOME=/usr/lib/jvm/java-11-openjdk-amd64

Then, apply the changes: Press **Ctrl + O** → press **Enter** to save.

Press **Ctrl + X** to exit.

source ~/.bashrc

# Step 4: Run Spark Shell

Start the Spark shell with the following command:

spark-shell

# Step 5: Run Sample Spark Logic

Paste this code inside the Spark shell to test basic RDD operations:

val numbers = List(-10, 0, 5, -3, 7)  
val rdd = sc.parallelize(numbers)  
val result = rdd.map { num =>  
 if (num > 0) s"$num is Positive"  
 else if (num < 0) s"$num is Negative"  
 else s"$num is Zero"  
}  
result.collect().foreach(println)