
⚙️ PART 1: Install and Configure Java, Scala, and Spark (on Windows)

This setup will:

- run Spark **natively on Windows**,
- let you **develop and run apps in IntelliJ IDEA**,
- optionally use **WSL 2** for Linux-style file access.

🧱 1. Install Java (JDK 17)

Spark 3.5.x works best with **Java 17**.

👉 In PowerShell (Run as Administrator)

```
winget install --id EclipseAdoptium.Temurin.17.JDK -e
```

Verify:

```
java -version
```

Expected:

```
openjdk version "17.0.x"
```

🧩 2. Install Scala (2.12.x)

Spark 3.5.1 requires Scala **2.12**.

```
choco install scala -y
```

(If you don't have Chocolatey yet, install it first:

```
Set-ExecutionPolicy Bypass -Scope Process -Force;
[System.Net.ServicePointManager]::SecurityProtocol =
[System.Net.ServicePointManager]::SecurityProtocol -bor 3072; iex
((New-Object
System.Net.WebClient).DownloadString('https://community.chocolatey.org
/install.ps1'))
```

Then re-run the Scala install.)

Verify:

```
scala -version
```

Output:

```
Scala code runner version 2.12.x
```

💾 3. Download Apache Spark (Pre-Built with Hadoop)

Download Spark 3.5.1 (Hadoop 3) pre-built package.

```
cd C:\
Invoke-WebRequest -Uri
"https://downloads.apache.org/spark/spark-3.5.1/spark-3.5.1-bin-hadoop
3.tgz" -OutFile "spark.tgz"
tar -xvzf spark.tgz
Rename-Item "spark-3.5.1-bin-hadoop3" "spark"
```

Now Spark is located at `C:\spark`.

⚙️ 4. Set Environment Variables

In PowerShell:

```
[Environment]::SetEnvironmentVariable("SPARK_HOME", "C:\spark",
"Machine")
[Environment]::SetEnvironmentVariable("HADOOP_HOME", "C:\spark",
"Machine")
[Environment]::SetEnvironmentVariable("JAVA_HOME", "C:\Program
Files\Eclipse Adoptium\jdk-17", "Machine")
[Environment]::SetEnvironmentVariable("PATH", $env:PATH +
";C:\spark\bin;C:\spark\sbin;C:\Program Files\Eclipse
Adoptium\jdk-17\bin", "Machine")
```

Close PowerShell and open a new one, then check:

```
spark-shell
```

✅ You should see:

```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.5.1
      /_/
```

Exit Spark shell:

```
:quit
```

🧰 PART 2: Setup in IntelliJ IDEA (Windows)

1️⃣ Install Plugins

Open IntelliJ IDEA →
File → Settings → Plugins → Marketplace
Install:

- ✅ Scala Plugin
- ✅ SBT (comes built-in with Community edition)

2️⃣ Create a New Spark Project

File → New → Project → Scala (SBT)

Set:

- Project name: `SparkHelloWorld`
- Scala version: `2.12.18`
- JDK: `17`

3️⃣ Edit `build.sbt`

Replace with this:

```
name := "SparkHelloWorld"
version := "1.0"
scalaVersion := "2.12.18"

libraryDependencies ++= Seq(
  "org.apache.spark" %% "spark-core" % "3.5.1",
  "org.apache.spark" %% "spark-sql"  % "3.5.1"
)
```

Click **Reload Project** when prompted.

4️⃣ Create a Sample Spark Application

Create file:
`src/main/scala/com/example/HelloSpark.scala`

```
package com.example

import org.apache.spark.sql.SparkSession

object HelloSpark {
  def main(args: Array[String]): Unit = {
    val spark = SparkSession.builder()
      .appName("HelloSparkApp")
      .master("local[*]")
      .getOrCreate()

    import spark.implicits._

    val data = Seq(("Asha", 30), ("Vikram", 25), ("Priya", 28))
    val df = data.toDF("name", "age")

    df.show()

    spark.stop()
  }
}
```

5️⃣ Run in IntelliJ

Click **Run** ▶️ on `HelloSpark`.

✅ Output:

```
+------+---+
|  name|age|
+------+---+
|  Asha| 30|
|Vikram| 25|
| Priya| 28|
+------+---+
```

💡 PART 3: Optional — Use WSL 2 with Spark

If you prefer to run Spark from **Ubuntu (WSL 2)** while developing in IntelliJ:

1. Place Spark folder in `/mnt/c/spark`

In WSL:

```
cd /mnt/c/spark
./bin/spark-shell
```

2.

You can also submit apps from WSL:

```
./bin/spark-submit --class com.example.HelloSpark
/mnt/c/Users/<YourUser>/IdeaProjects/SparkHelloWorld/target/scala-2.12
/sparkhelloworld_2.12-1.0.jar
```

3.

This way, Spark runs in a Linux-like environment, but your development stays in Windows.

🚀 PART 4: Test GitHub Repository

Clone this **beginner-friendly Spark + Scala** repo:
👉 https://github.com/awesome-spark/spark-scala-hello

In PowerShell:

```
cd C:\
git clone https://github.com/awesome-spark/spark-scala-hello.git
cd spark-scala-hello
```

Then open the project in IntelliJ → **Reload Project** → **Run**.

✅ Summary Checklist

Step	Task	Tool
1	Install Java 17	PowerShell
2	Install Scala 2.12	PowerShell / Chocolatey
3	Download Spark 3.5.1	PowerShell
4	Set Environment Variables	PowerShell
5	Test Spark shell	PowerShell
6	Create Spark Project in IntelliJ	IntelliJ
7	Add build.sbt dependencies	IntelliJ
8	Run HelloSpark.scala	IntelliJ
9	(Optional) Run Spark in WSL	WSL Terminal
