PySpark Setup on Local

For macOS

1. **Install Homebrew** (if you haven't already): Open your terminal and run:

```
/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"
```

2. **Install Java**: PySpark requires Java. You can install it using Homebrew:

```
brew install openjdk@11
```

After installation, you may need to set the JAVA_HOME environment variable. Add this to your `~/.bash_profile` or `~/.zshrc`:

```
export JAVA_HOME="$(brew --prefix openjdk@11)"
```

3. **Install Apache Spark**: Use Homebrew to install Spark:

```
brew install apache-spark
```

4. **Install PySpark**: You can install PySpark via pip:

```
pip install pyspark
```

5. **Run PySpark**: You can run PySpark in the terminal using:

```
pyspark
```

For Windows

1. **Install Java**: Download and install the Java Development Kit (JDK) from Oracle's website or use OpenJDK. Make sure to set the `JAVA_HOME` environment variable to the JDK installation path.

2. **Install Spark**:

   - Download the latest version of Apache Spark from Spark's official website.

   - Extract the downloaded archive to a directory of your choice.

3. **Set Environment Variables**:
   - Add the Spark `bin` directory to your PATH environment variable.
   - Set the `SPARK_HOME` environment variable to the Spark installation directory.

4. **Install PySpark**: Open a command prompt and install PySpark using pip:

```
pip install pyspark
```

5. **Run PySpark**: Open a command prompt and run:

```
pyspark
```

Testing Your Setup

Once you have everything installed, you can create a simple PySpark script to test it. Create a file called `test_spark.py`:

```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("Test Spark") \
    .getOrCreate()

data = [("Alice", 1), ("Bob", 2)]
df = spark.createDataFrame(data, ["Name", "Value"])
df.show()

spark.stop()
```

Run this script using:

```
python test_spark.py
```

Notes

- Running Spark locally will use your local resources, so performance may vary compared to a cloud setup.
- Make sure to have sufficient memory and CPU resources available on your local machine for testing larger datasets.
- If you encounter issues, refer to the Spark logs for troubleshooting.