

Apache Spark Installation Guide – macOS

Pre-requisites

- Java Version: JDK 17.0.17
- Python Version: 3.10.0
- Spark Version: 3.5.7
- Note: Hadoop winutils not required on macOS

Installation Steps

Step 1

Download and install Java 17 for macOS from:

<https://www.oracle.com/java/technologies/downloads/#java17-mac>

Verify installation: `java -version`

Step 2

Download and install Python 3.10 for macOS from:

<https://www.python.org/ftp/python/3.10.0/python-3.10.0-macosx10.9.pkg>

Verify installation: `python3 --version`

Step 3

Download Spark 3.5.7 prebuilt for Hadoop 3:

<https://www.apache.org/dyn/closer.lua/spark/spark-3.5.7/spark-3.5.7-bin-hadoop3.tgz>

Extract the archive: `tar -xzf spark-3.5.7-bin-hadoop3.tgz`

Rename folder to 'spark' and move to /usr/local/: `sudo mv spark-3.5.7-bin-hadoop3 /usr/local/spark`

Step 4

Set environment variables in your shell profile (~/.zshrc or ~/.bash_profile):

```
export SPARK_HOME=/usr/local/spark
```

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

```
export JAVA_HOME=$(/usr/libexec/java_home -v 17)
```

```
export PYSPARK_PYTHON=$(which python3)
```

```
export PYSPARK_DRIVER_PYTHON=$(which python3)
```

Reload shell: `source ~/.zshrc` or `source ~/.bash_profile`

Step 5

Verify installation by running:

```
java -version  
python3 --version  
spark-shell  
pyspark
```

Step 6

Test a small PySpark job:

```
from pyspark.sql import SparkSession  
spark = SparkSession.builder.appName('TestSparkMac').getOrCreate()  
data = [('John', 30), ('Alice', 25)]  
df = spark.createDataFrame(data, ['name', 'age'])  
df.show()  
spark.stop()
```