

Spark mit Windows 10

<https://phoenixnap.com/kb/install-spark-on-windows-10>

Anleitung zur Installation von Apache Spark auf Windows 10. Dies bildet die Basis für den Einfachen Einsatz und Nutzung durch Sprachen wie Kotlin oder C# (Auch das geht, wie man hier sieht: <https://docs.microsoft.com/de-de/dotnet/spark/what-is-apache-spark-dotnet>)

Für die Nutzung von Spark bedarf es mindestens vier Vorgaben:

1. Java muss in einer neueren Version installiert sein
2. Python muss in einer neueren Version installiert sein
3. Apache Spark selbst muss installiert sein
4. Zudem muss die Zusammenarbeit mit Hadoop berücksichtigt werden

Hinweis: In den Abbildungen tauchen zum Teil unterschiedliche Versionen der Software auf. Dies ist darauf zurückzuführen, dass ich an zwei Rechnern jeweils unterschiedliche Versionen installiert habe und die Screenshots während dieser beiden Installationen entstanden.

Java

Mit Hilfe des Befehls `java -version` kann die aktuell installierte Version überprüft werden. Gleichzeitig ist so sichergestellt, dass die Systemvariablen für Java korrekt eingestellt wurden:

```
PS C:\> java -version
java version "1.8.0_261"
Java(TM) SE Runtime Environment (build 1.8.0_261-b12)
Java HotSpot(TM) Client VM (build 25.261-b12, mixed mode)
PS C:\>
```

Ist Java auf dem System installiert erscheint eine Meldung ähnlich der oben zu sehenden. Hier im Beispiel ist die Version 1.8 installiert, welche aktuell ist.

Sollte dies nicht der Fall sein, so kann unter <https://java.com/en/download/> eine aktuelle Version installiert werden. Hier findet man auch die entsprechenden Schritte zur Installation.

Python

Mit Hilfe des Befehls `python --version` kann die Version der installierten Python Version abgerufen werden. Auch hier signalisiert der erfolgreiche Aufruf, dass die Umgebungsvariablen gesetzt wurden.

```
PS C:\> python --version
Python 3.9.6
PS C:\> |
```

In diesem Fall ist die aktuelle Version 3.9.6 vorhanden. Für den Fall, dass Python nicht installiert ist, können die notwendigen Installationsdatei und Anleitung auf der Seite von <https://www.python.org> abgerufen werden. Der Download selbst befindet sich beispielsweise hier: <https://www.python.org/downloads/release/python-396/> (hier für die Version 3.9.6).

Spark

<https://spark.apache.org/downloads.html>

Die notwendigen Dateien finden sich an der oben genannten Adresse. Hier kann unter anderen die Release Version sowie der Typ eingestellt werden. An dieser Stelle wird die Auswahl wie nachfolgend zu sehen getroffen:

Download Apache Spark™

1. Choose a Spark release:
2. Choose a package type:
3. Download Spark: [spark-3.0.3-bin-hadoop3.2.tgz](#)
4. Verify this release using the 3.0.3 [signatures](#), [checksums](#) and [project release KEYS](#).

Über den an Punkt 3 gezeigten Link gelangt man zu einer Seite mit verschiedenen Mirrors, wobei der oberste in der Regel eine gute Wahl darstellt.

We suggest the following mirror site for your download:

<https://mirror.checkdomain.de/apache/spark/spark-3.0.3/spark-3.0.3-bin-hadoop3.2.tgz>

Other mirror sites are suggested below.

Nach dem Download kann die heruntergeladene Datei mit Hilfe der Checksumme verifiziert werden. Hierzu folgt man den unter Punkt 4 angegebenen Link zur Checksumme:

← → ↻ 🔒 downloads.apache.org/spark/spark-3.0.3/spark-3.0.3-bin-hadoop3.2.tgz.sha512

Apps Google Kalender Virtuelle Fachhochs... WEB.DE - E-Mail-A...

spark-3.0.3-bin-hadoop3.2.tgz: 22FC9A60 42769D13 C01F3B07 F0DFAE9D 5D1E4082
44870833 C5E703EB 43669CE1 4C875F7C EC47A07F
1CF2EE3E 65DFB136 39776448 CD686481 CCC0E071
5D2760D5




Mit dem Befehl **certutil** kann nun die Checksumme aus der heruntergeladenen Datei generiert und mit der angezeigten verglichen werden. Hierbei sollte darauf geachtet werden, dass die Versionen zueinander passen, da es leicht zu Verwechslungen kommen kann.

```
PS C:\> certutil -hashfile c:\users\username\Downloads\spark-3.0.3-bin-hadoop3.2.tgz SHA512
CertUtil: -hashfile-Befehl ist fehlgeschlagen: 0x80070002 (WIN32: 2 ERROR_FILE_NOT_FOUND)
CertUtil: Das System kann die angegebene Datei nicht finden.
PS C:\> certutil -hashfile C:\Users\ckitte\Downloads\spark-3.0.3-bin-hadoop3.2.tgz SHA512
SHA512-Hash von C:\Users\ckitte\Downloads\spark-3.0.3-bin-hadoop3.2.tgz:

22fc9a6042769d13c01f3b07f0dfae9d5d1e408244870833c5e703eb43669ce14c875f7cec47a07f1cf2ee3e65dfb1363977644bcd686481ccc0e0715d2760d5
CertUtil: -hashfile-Befehl wurde erfolgreich ausgeführt.
PS C:\> |
```

Die Installation von Spark ist eigentlich ein Entpacken der heruntergeladenen Datei in ein frei zu bestimmendes Verzeichnis. Für das Entpacken kann das normale und in Windows vorhandene ZIP Tool verwendet werden.

Nach dem Entpacken der tgz Datei, existiert eine neue Datei mit der Endung tar. Hierbei handelt es sich um ein gezipptes Verzeichnis. Nachdem diese Datei ebenfalls entpackt wurde, liegt das eigentliche Verzeichnis vor:

<input type="checkbox"/> Name	Änderungsdatum	Typ	Größe
 spark-3.0.3-bin-hadoop3.2	17.06.2021 06:51	Dateiordner	
 spark-3.0.3-bin-hadoop3.2.tar	07.07.2021 15:32	TAR-Datei	245.320 KB
 spark-3.0.3-bin-hadoop3.2.tgz	07.07.2021 15:32	TGZ-Datei	219.496 KB

In dem angelegten Ordner befinden sich alle zu Spark gehörenden Dateien, welche nun verfügbar sind.

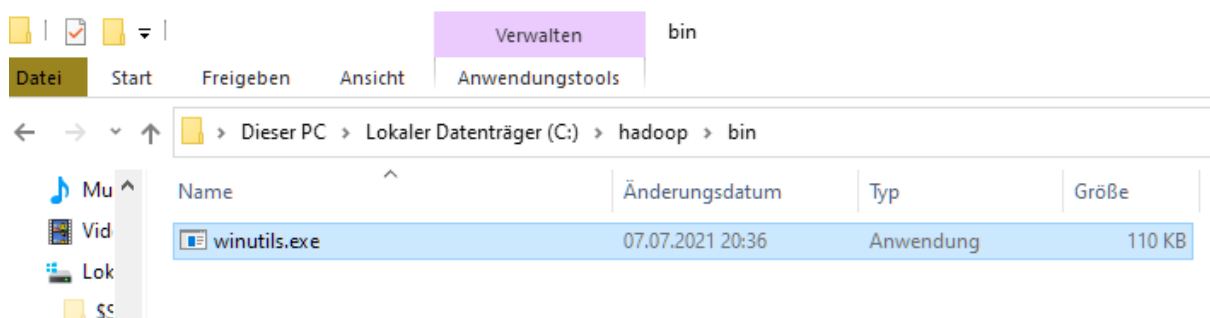
Spark baut auf Hadoop auf. die hier verwendete Version 3.2 wurde beim Download (siehe oben) ausgewählt. Zum Betrieb unter Windows muss jedoch keine komplette Hadoop Umgebung realisiert werden. Allerdings wird unter Windows ein kleines Tool benötigt, da Hadoop Probleme mit dem Dateiverzeichnis von Windows hat.

Das Tool kann unter <https://github.com/cdarlint/winutils> passend für die gewählte Version herunter geladen werden.

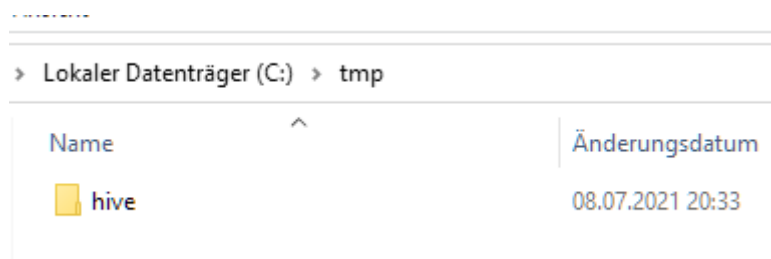
master winutils / hadoop-3.2.1 / bin /

cdarlint add 321 winutils ...

Hierzu wird in das passende Verzeichnis auf GitHub gegangen und die Datei winutils.exe heruntergeladen. Diese wird anschließend in einem Verzeichnis c:\hadoop\bin hinein kopiert.



Weiter benötigt Hadoop nach Aussage einiger Quellen für das korrekte Funktionieren ein temporäres Verzeichnis unter c:\tmp\hive zur Speicherung von Zwischenwerte. Diese Aussage konnte jedoch nicht sicher geprüft werden.



Das angelegte Verzeichnis muss für winutil verfügbar sein. Dies erreicht man durch das Ausführen des folgenden Kommandos:

```
PS C:\Users\Christian Kitte> winutils chmod 777 c:\tmp\hive
PS C:\Users\Christian Kitte> |
```

Abschließend müssen im nächsten Schritt noch einige Pfade angepasst werden, damit alle Komponenten zusammenarbeiten können. Die folgenden Umgebungsvariablen müssen hierbei festgelegt, sowie die Pfadvariablen entsprechend erweitert werden:

SPARK_HOME

c:\spark\spark-3.1.3-bin-hadoop3.2

Path ergänzen mit %SPARK_HOME%\bin

HADOOP_HOME

C:\hadoop

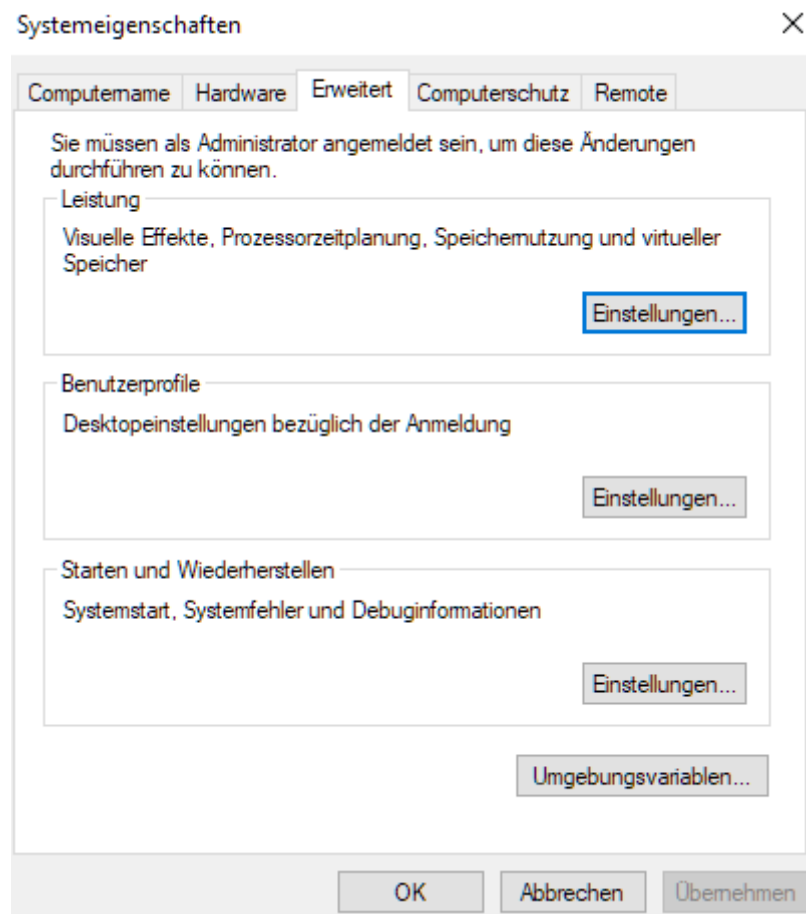
Path ergänzen mit %HADOOP_HOME%\bin

JAVA_HOME

c:\Program Files (x86)\Java\jre1.8.0_291 (jre reicht)

keine Pfad

Für die Umsetzung wechselt man über **System** in **Systemeigenschaften** und hier in die Bearbeitung der **Umgebungsvariablen** (Button unten rechts auf dem Reiter "Erweitert"):



Im Anschluss sollten alle Umgebungsvariablen verfügbar sein (hier in Powershell):

```
PS C:\> $env:SPARK_HOME
c:\spark\spark-3.1.2-bin-hadoop3.2
PS C:\> $env:HADOOP_HOME
c:\hadoop
PS C:\> $env:JAVA_HOME
C:\Program Files (x86)\Java\jre1.8.0_291
PS C:\> |
```

und hier in der normalen Eingabeaufforderung:

```
C:\Users\Christian Kitten>echo %SPARK_HOME%
c:\spark\spark-3.1.2-bin-hadoop3.2

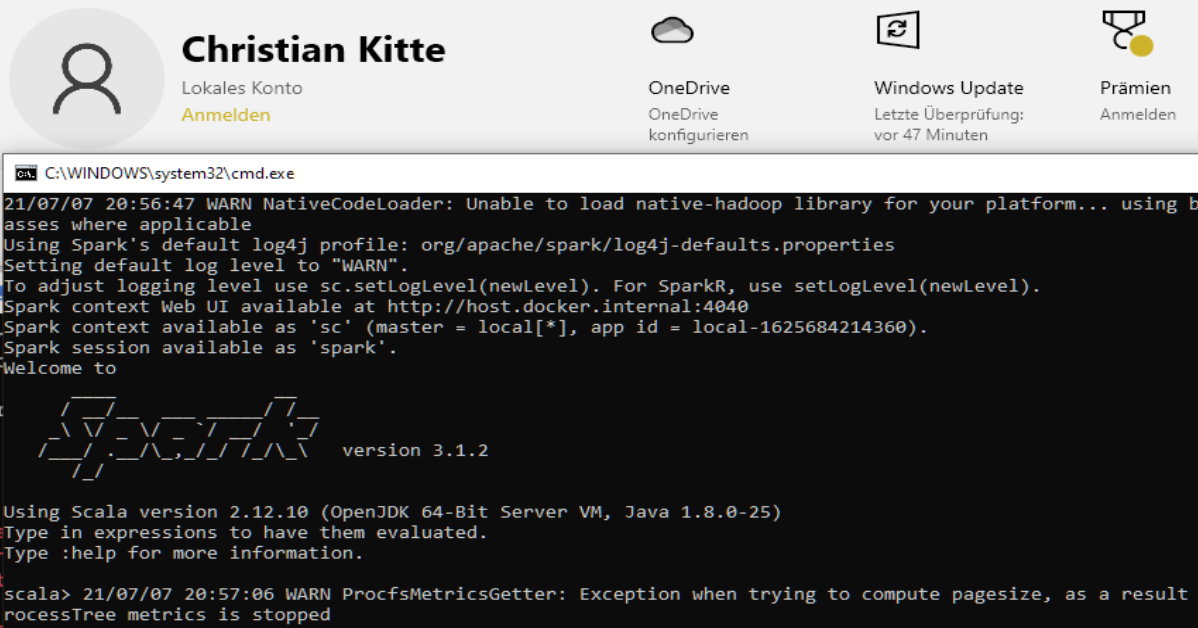
C:\Users\Christian Kitten>echo %HADOOP_HOME%
c:\hadoop

C:\Users\Christian Kitten>echo %JAVA_HOME%
C:\Program Files (x86)\Java\jre1.8.0_291
```

Somit ist nunmehr die Installation abgeschlossen. mit dem Befehl **spark-shell** lässt sich die Konsole von Spark aufrufen;

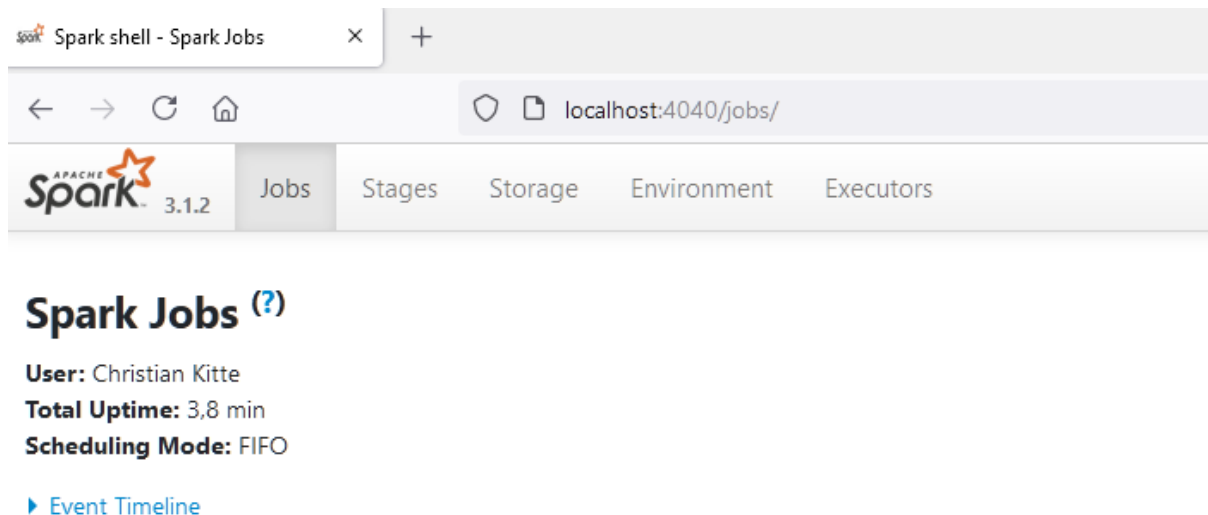
```
PS C:\> C:\spark\spark-3.1.2-bin-hadoop3.2\bin\spark-shell
PS C:\> |
```

Nach der Bestätigung öffnet sich eine neue Konsole und Spark startet mit einer Reihe von Ausgaben.



The screenshot shows a Windows desktop environment. At the top, there is a user profile for 'Christian Kitten' with a 'Lokales Konto' (Local Account) and an 'Anmelden' (Sign in) button. To the right of the profile are four utility icons: OneDrive (OneDrive konfigurieren), Windows Update (Letzte Überprüfung: vor 47 Minuten), and Prämien (Anmelden). Below the desktop background, a terminal window titled 'C:\WINDOWS\system32\cmd.exe' is open. The terminal displays the output of the 'spark-shell' command. It shows a warning from 'NativeCodeLoader' about not being able to load the native-hadoop library, followed by Spark's default log4j profile settings. It then displays the Spark context Web UI available at 'http://host.docker.internal:4040' and the Spark session available as 'spark'. A 'Welcome to' message is followed by the Spark logo and 'version 3.1.2'. Below this, it shows 'Using Scala version 2.12.10 (OpenJDK 64-Bit Server VM, Java 1.8.0-25)' and prompts the user to 'Type in expressions to have them evaluated.' and 'Type :help for more information.' The terminal also shows a warning from 'ProcsMetricsGetter' about an exception when trying to compute pagesize, as a result of 'processTree metrics is stopped'.

Gleichzeitig ist die shell Web UI von Spark über den Browser am standardmäßigen Port 4040 erreichbar:



Spark mit dem WSL

WSL steht für Windows Subsystem für Linux und beinhaltet eine komplette Umgebung für Linux. Dies ermöglicht z.B. den einfachen und nativen Umgang mit Docker Container für Linux ohne Umwege.

Das WSL kann problemlos erweitert werden. So habe ich beispielsweise ein komplettes Ubuntu System eingerichtet:

The image shows a terminal window with the title 'ckitte@Baseking: /mnt/c'. The prompt is 'ckitte@Baseking:/mnt/c\$'. The command 'lsb_release -d' has been executed, and the output is 'Description: Ubuntu 20.04.2 LTS'. The prompt is now 'ckitte@Baseking:/mnt/c\$ |'.

Im folgenden habe ich auf diesem System zunächst Java, dann Python sowie abschließend PySpark separat installiert:

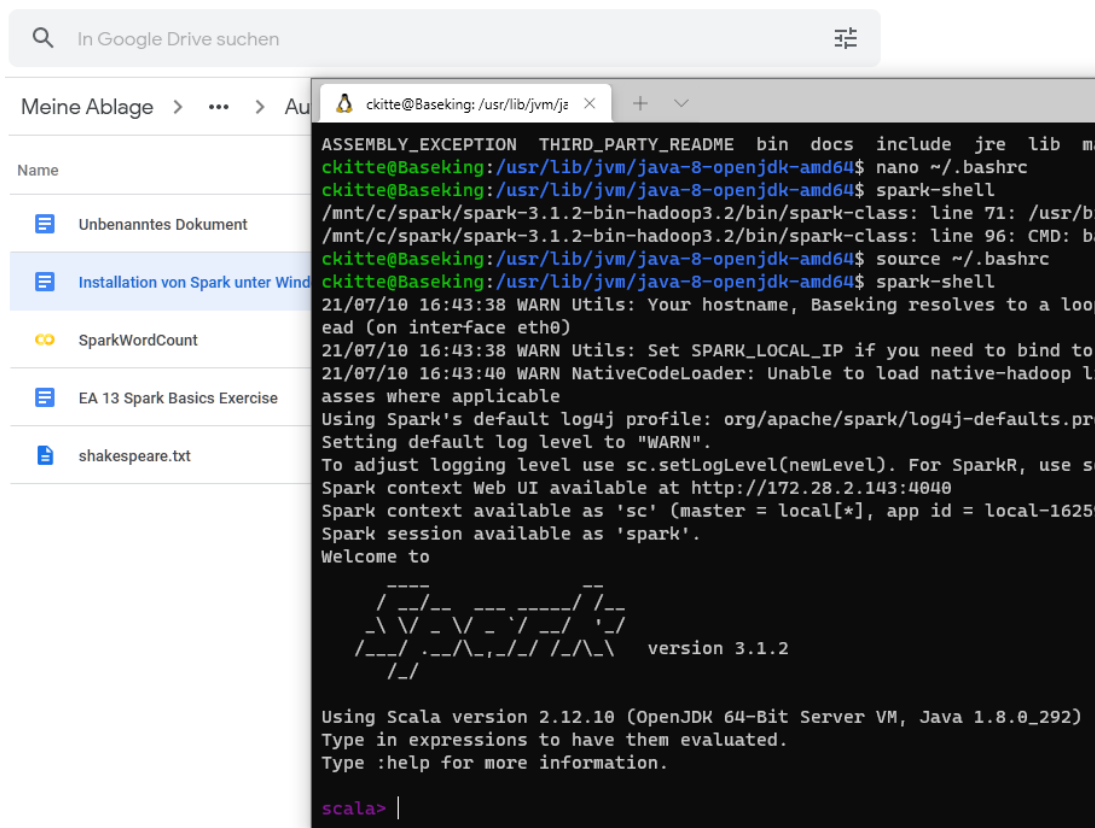
```
sudo apt-get -y install openjdk-8-jdk-headless  
sudo apt install python3  
sudo apt install pyspark
```

Obwohl die Pfade der Anleitung entsprechend gesetzt waren, kam es wie zuvor bei Windows selbst und entgegen der Vorhersage der Anleitung, zu Fehlern beim Aufruf. Entweder wurden Java Klassen nicht gefunden, oder Skripte wie spark-submit, welcher definitiv vorhanden waren.

(Exkurs: Pfade können unter Linux mit Hilfe der Datei **bashrc** definiert werden. Ich verwende hier gerne **nano** als einfachen Editor. Mit **nano ~/.bashrc** wird die Datei geöffnet. Nach dem Speichern und schließen müssen die Änderungen bekannt gemacht werden mit **source ~/.bashrc**.)

```
SPARK_HOME=/mnt/c/spark/spark-3.1.2-bin-hadoop3.2
export PATH=$SPARK_HOME/bin:$PATH
JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PATH=$JAVA_HOME/bin:$PATH
```

SPARK_HOME zeigt hierbei auf das auch für Windows verwendete Verzeichnis ohne den Anhang **bin**. Hierbei verwende ich den **vollen Pfad**. Ebenso verwende ich für **JAVA_HOME** den **vollen Pfad**. Als Ergebnis konnte ich Spark unter Windows innerhalb des Windows Subsystems für Linux sowohl für Scala:



als auch für Python:

```
ckitte@Baseking:/usr/lib/jvm/java-8-openjdk-amd64$ pyspark
Python 3.8.10 (default, Jun  2 2021, 10:49:15)
[GCC 9.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
21/07/10 16:46:09 WARN Utils: Your hostname, Baseking resolves to a loopback address
ead (on interface eth0)
21/07/10 16:46:09 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another add
21/07/10 16:46:11 WARN NativeCodeLoader: Unable to load native-hadoop library for yo
asses where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(ne
21/07/10 16:46:14 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempt
Welcome to

  _--_
 /  _/ _-- _-- _-- _-- _/ _/ _/
 _\ \ / _ \ / _ \ / _ \ / _ \
/_ _ / _ _ \ _ \ _ \ / _ \ _ \   version 3.1.2
 / _/

Using Python version 3.8.10 (default, Jun  2 2021 10:49:15)
Spark context Web UI available at http://172.28.2.143:4041
Spark context available as 'sc' (master = local[*], app id = local-1625928374353).
SparkSession available as 'spark'.
>>> |
```

starten.