

Установка one-node Spark на Ubuntu

- Установка JDK

sudo apt update

```
default@Main:~$ sudo apt update
[sudo] password for default:
Hit:1 http://archive.ubuntu.com/ubuntu jammy InRelease
Hit:2 http://archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:3 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:4 http://security.ubuntu.com/ubuntu jammy-security InRelease
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
All packages are up to date.
```

sudo apt install openjdk-11-jdk

начало установки:

```
default@Main:~$ sudo apt install openjdk-11-jdk
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  alsa-topology-conf alsa-ucm-conf ca-certificates-java fonts-dejavu-extra java-common libasound2 libasound2-data
  libatk-wrapper-java libatk-wrapper-java-jni libfontenc1 libgif7 libice-dev libice6 libnspr4 libnss3 libpcsclite1
  libpthread-stubs0-dev libsm-dev libsm6 libx11-dev libxau-dev libxaw7 libxcb-shape0 libxcb1-dev libxdmcp-dev libxft2
  libxkbfile1 libxmu6 libxpm4 libxt-dev libxt6 libxv1 libxxf86dga1 openjdk-11-jdk-headless openjdk-11-jre
  openjdk-11-jre-headless x11-utils x11proto-dev xorg-sgml-doctools xtrans-dev
Suggested packages:
  default-jre libasound2-plugins alsa-utils libice-doc pcscd libsm-doc libx11-doc libxcb-doc libxt-doc openjdk-11-demo
  openjdk-11-source visualvm libnss-mdns fonts-ipafont-gothic fonts-ipafont-mincho fonts-wqy-microhei
  | fonts-wqy-zenhei fonts-indic mesa-utils
The following NEW packages will be installed:
  alsa-topology-conf alsa-ucm-conf ca-certificates-java fonts-dejavu-extra java-common libasound2 libasound2-data
  libatk-wrapper-java libatk-wrapper-java-jni libfontenc1 libgif7 libice-dev libice6 libnspr4 libnss3 libpcsclite1
  libpthread-stubs0-dev libsm-dev libsm6 libx11-dev libxau-dev libxaw7 libxcb-shape0 libxcb1-dev libxdmcp-dev libxft2
  libxkbfile1 libxmu6 libxpm4 libxt-dev libxt6 libxv1 libxxf86dga1 openjdk-11-jdk openjdk-11-jdk-headless
  openjdk-11-jre openjdk-11-jre-headless x11-utils x11proto-dev xorg-sgml-doctools xtrans-dev
0 upgraded, 41 newly installed, 0 to remove and 0 not upgraded.
Need to get 125 MB of archives.
After this operation, 285 MB of additional disk space will be used.
Do you want to continue? [Y/n] y
```

окончание установки:

```
Adding debian:HiPKI_Root_CA_-_G1.pem
Adding debian:Baltimore_CyberTrust_Root.pem
Adding debian:Microsoft_ECC_Root_Certificate_Authority_2017.pem
Adding debian:TWCA_Root_Certification_Authority.pem
Adding debian:Trustwave_Global_ECC_P256_Certification_Authority.pem
Adding debian:Telia_Root_CA_v2.pem
Adding debian:GTS_Root_R1.pem
Adding debian:D-TRUST_Root_Class_3_CA_2_2009.pem
Adding debian:emSign_Root_CA_-_G1.pem
Adding debian:Buypass_Class_3_Root_CA.pem
Adding debian:SwissSign_Silver_CA_-_G2.pem
Adding debian:GlobalSign_Root_CA.pem
Adding debian:GDCA_TrustAUTH_R5_ROOT.pem
Adding debian:certSIGN_Root_CA_G2.pem
Adding debian:CFCA_EV_ROOT.pem
Adding debian:CommScope_Public_Trust_RSA_Root-01.pem
Adding debian:GlobalSign_Root_CA_-_R6.pem
done.
Processing triggers for fontconfig (2.13.1-4.2ubuntu5) ...
Processing triggers for hicolor-icon-theme (0.17-2) ...
Processing triggers for libc-bin (2.35-0ubuntu3.9) ...
Processing triggers for man-db (2.10.2-1) ...
Processing triggers for ca-certificates (20240203~22.04.1) ...
Updating certificates in /etc/ssl/certs...
0 added, 0 removed; done.
Running hooks in /etc/ca-certificates/update.d...

done.
done.
default@Main:~$
```

Проверка установки:

```
default@Main:~$ java -version
openjdk version "11.0.26" 2025-01-21
OpenJDK Runtime Environment (build 11.0.26+4-post-Ubuntu-1ubuntu122.04)
OpenJDK 64-Bit Server VM (build 11.0.26+4-post-Ubuntu-1ubuntu122.04, mixed mode, sharing)
default@Main:~$
```

- Настройка переменной \$JAVA_HOME

```
echo "export JAVA_HOME=$(dirname $(dirname $(readlink -f $(which java))))" >> ~/.bashrc
source ~/.bashrc
```

Проверка установки:

```
default@Main:~$ echo $JAVA_HOME
/usr/lib/jvm/java-11-openjdk-amd64
```

- Установка Spark

```
wget https://archive.apache.org/dist/spark/spark-3.3.1/spark-3.3.1-bin-hadoop3.tgz
tar -xvzf spark-3.3.1-bin-hadoop3.tgz
sudo mv spark-3.3.1-bin-hadoop3 /opt/spark
```

```
default@Main:~$ wget https://archive.apache.org/dist/spark/spark-3.3.1/spark-3.3.1-bin-hadoop3.tgz
--2025-05-04 15:16:19-- https://archive.apache.org/dist/spark/spark-3.3.1/spark-3.3.1-bin-hadoop3.tgz
Resolving archive.apache.org (archive.apache.org)... 65.108.204.189, 2a01:4f9:1a:a084::2
Connecting to archive.apache.org (archive.apache.org)|65.108.204.189|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 299350810 (285M) [application/x-gzip]
Saving to: 'spark-3.3.1-bin-hadoop3.tgz'

spark-3.3.1-bin-hadoop3.tgz      100%[=====] 285.48M  24.8MB/s   in 12s

2025-05-04 15:16:33 (23.4 MB/s) - 'spark-3.3.1-bin-hadoop3.tgz' saved [299350810/299350810]
```

```
default@Main:~$ tar -xvzf spark-3.3.1-bin-hadoop3.tgz
spark-3.3.1-bin-hadoop3/
spark-3.3.1-bin-hadoop3/LICENSE
spark-3.3.1-bin-hadoop3/NOTICE
spark-3.3.1-bin-hadoop3/R/
spark-3.3.1-bin-hadoop3/R/lib/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/DESCRIPTION
spark-3.3.1-bin-hadoop3/R/lib/SparkR/INDEX
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/
```

```
default@Main:~$ sudo mv spark-3.3.1-bin-hadoop3 /opt/spark
```

- Настройка переменной \$SPARK_HOME

```
echo "export SPARK_HOME=/opt/spark" >> ~/.bashrc
echo "export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin" >> ~/.bashrc
source ~/.bashrc
```

```
default@Main:~$ echo "export SPARK_HOME=/opt/spark" >> ~/.bashrc
default@Main:~$ nano ~/.bashrc
default@Main:~$ echo "export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin" >> ~/.bashrc
default@Main:~$ nano ~/.bashrc
default@Main:~$ source ~/.bashrc
default@Main:~$
```

Проверка:

```
default@Main:~$ echo $SPARK_HOME
/opt/spark
default@Main:~$
```

- Запуск примера

```
$SPARK_HOME/bin/run-example SparkPi 10
```

```
25/05/04 16:13:04 INFO DAGScheduler: Job 0 finished: reduce at SparkPi.scala:38, took 0.422816 s
Pi is roughly 3.14005914005914
```

Установка кластера Spark на Docker

- Запуск docker-compose

`docker-compose up -d`

```
PS D:\YandexDisk\Education\spark-otus\homework\hw-01> docker-compose up -d
[+] Running 5/5
✔ Container spark-worker-03 Started 23.8s
✔ Container spark-worker-01 Started 23.9s
✔ Container hw-01-almond-1 Started 24.0s
✔ Container spark-worker-02 Started 23.9s
✔ Container spark-master Started 12.0s
```

- Копирование файла `simple_app.py` в каталог «scripts»
- Запуск `spark` – приложения из контейнера

`docker-compose exec spark-master /opt/bitnami/spark/bin/spark-submit --master spark://spark-master:7077 /scripts/simple_app.py`

```
+-----+-----+
|Language| Users|
+-----+-----+
|   Java | 20000|
| Python|100000|
|   Scala|  3000|
+-----+-----+
```

```
+-----+-----+
|Language|count|
+-----+-----+
|   Java |    1|
|   Scala|    1|
| Python|    1|
+-----+-----+
```

```
default@Main:~$ docker exec -u 1001 -it spark-master \
/opt/bitnami/spark/bin/spark-submit \
--master spark://spark-master:7077 \
--class org.apache.spark.examples.SparkPi \
/opt/bitnami/spark/examples/jars/spark-examples_2.12-3.5.5.jar \
10
```

```
25/05/04 20:36:14 INFO DAGScheduler: Job 0 finished: reduce at SparkPi.scala:38, took 2.221446 s
Pi is roughly 3.138775138775139
25/05/04 20:36:14 INFO SparkContext: SparkContext is stopping with exitCode 0.
25/05/04 20:36:14 INFO SparkUI: Stopped Spark web UI at http://spark-master:4040
25/05/04 20:36:14 INFO StandaloneSchedulerBackend: Shutting down all executors
25/05/04 20:36:14 INFO StandaloneSchedulerBackend$StandaloneDriverEndpoint: Asking each executor to shut down
25/05/04 20:36:14 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/05/04 20:36:14 INFO MemoryStore: MemoryStore cleared
25/05/04 20:36:14 INFO BlockManager: BlockManager stopped
25/05/04 20:36:14 INFO BlockManagerMaster: BlockManagerMaster stopped
25/05/04 20:36:14 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/05/04 20:36:14 INFO SparkContext: Successfully stopped SparkContext
25/05/04 20:36:14 INFO ShutdownHookManager: Shutdown hook called
25/05/04 20:36:14 INFO ShutdownHookManager: Deleting directory /tmp/spark-e105be80-7cd1-4913-af16-8d756bba1e1d
25/05/04 20:36:14 INFO ShutdownHookManager: Deleting directory /tmp/spark-da7c20e7-ea0d-4d76-bb0a-366f1c2aa240
default@Main:~$
```

Completed Applications (3)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20250504203611-0002	Spark Pi	48	1024.0 MiB		2025/05/04 20:36:11	spark	FINISHED	3 s
app-20250504195919-0001	SimplePySparkApp	48	1024.0 MiB		2025/05/04 19:59:19	spark	FINISHED	8 s
app-20250504195315-0000	SimplePySparkApp	48	1024.0 MiB		2025/05/04 19:53:15	spark	FINISHED	10 s