

Spark

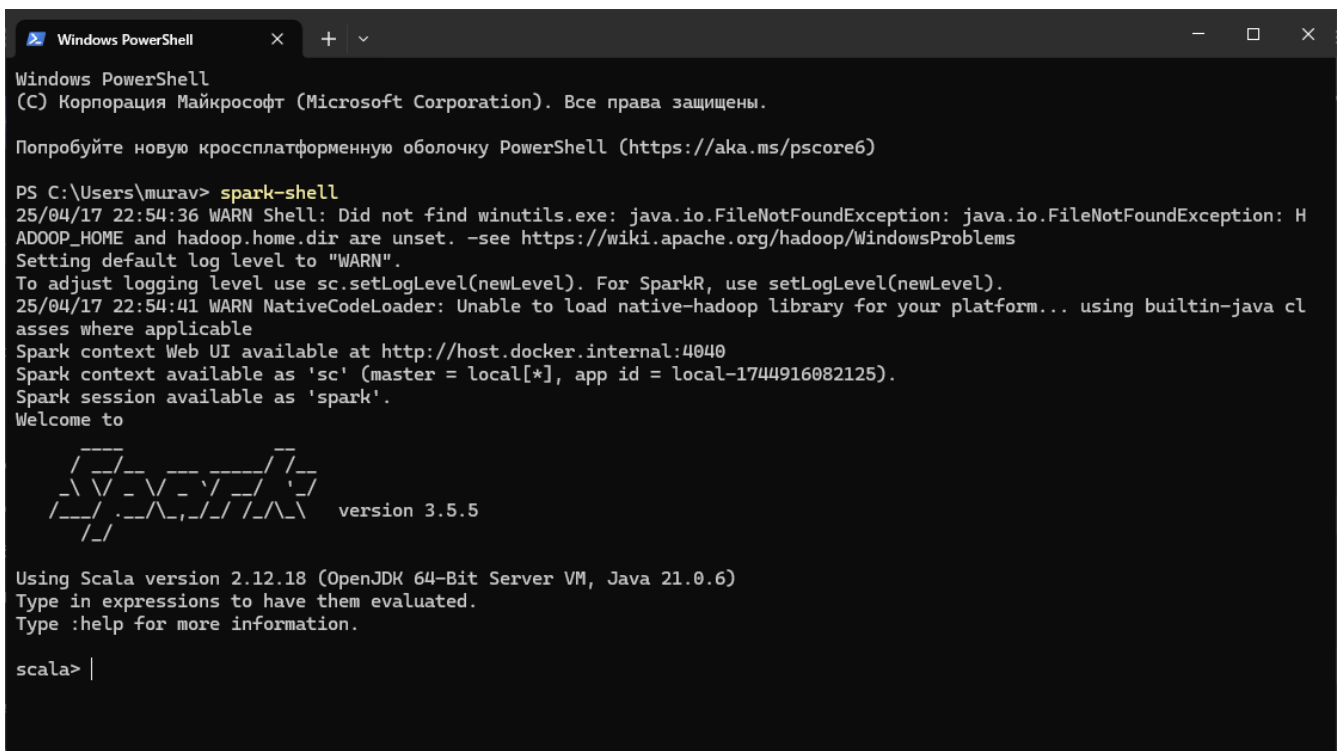
- 1) Установите Spark локально на операционную систему.
- 2) Запустите интерпретатор Scala/Python и выполните простой код (например, подсчет количества слов в тексте)

```
val text = "Hello world"

val wordCounts = text.split(" ")
  .groupBy(identity)
  .mapValues(_.length)

println(wordCounts)
```

- 3) При помощи встроенных средств выведите информацию о Spark в консоль



```
Windows PowerShell
(C) Корпорация Майкрософт (Microsoft Corporation). Все права защищены.

Попробуйте новую кроссплатформенную оболочку PowerShell (https://aka.ms/pscore6)

PS C:\Users\murav> spark-shell
25/04/17 22:54:36 WARN Shell: Did not find winutils.exe: java.io.FileNotFoundException: java.io.FileNotFoundException: H
ADOOP_HOME and hadoop.home.dir are unset. -see https://wiki.apache.org/hadoop/WindowsProblems
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/04/17 22:54:41 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Spark context Web UI available at http://host.docker.internal:4040
Spark context available as 'sc' (master = local[*], app id = local-1744916082125).
Spark session available as 'spark'.
Welcome to

  ____      _
 / ___|  _ \| | | |
 \___ \| |_) | |_| |
  ___) | |_) | | | |
 |___) |___|_|_|_|_|
               version 3.5.5

Using Scala version 2.12.18 (OpenJDK 64-Bit Server VM, Java 21.0.6)
Type in expressions to have them evaluated.
Type :help for more information.

scala> |
```

- 4) Выведите web-страницу с информацией о Spark и его текущей загрузке

Executors

[Show Additional Metrics](#)

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	0.0 B / 434.4 MiB	0.0 B	8	0	0	0	0	6.6 min (0.1 s)	0.0 B	0.0 B	0.0 B	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	0.0 B / 434.4 MiB	0.0 B	8	0	0	0	0	6.6 min (0.1 s)	0.0 B	0.0 B	0.0 B	0

Executors

Show 20 entries


Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump	Heap Histogram	Add Time	Remove Time
driver	host.docker.internal:50481	Active	0	0.0 B / 434.4 MiB	0.0 B	8	0	0	0	0	6.6 min (0.1 s)	0.0 B	0.0 B	0.0 B	Thread Dump	Heap Histogram	2025-04-17 22:54:42	-

Showing 1 to 1 of 1 entries

Previous 1 Next

5) Установите Spark на кластер Hadoop, созданный в предыдущей практической работе



Spark Master at spark://spark-master:7077

URL: spark://spark-master:7077

Alive Workers: 2

Cores in use: 4 Total, 0 Used

Memory in use: 4.0 GiB Total, 0.0 B Used

Resources in use:

Applications: 0 Running, 0 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20250417191959-172.19.0.11-39707	172.19.0.11:39707	ALIVE	2 (0 Used)	2.0 GiB (0.0 B Used)	
worker-20250417192000-172.19.0.12-38353	172.19.0.12:38353	ALIVE	2 (0 Used)	2.0 GiB (0.0 B Used)	

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

6) При помощи встроенных средств выведите информацию о Spark в консоль

```
spark-submit --version
```

```
I have no name!@spark-master:/opt/bitnami/spark$ spark-submit --version
Welcome to

  ____      __
 / __ \__  _/  ____/  __/
_\" \ / _\" \ / _\" \ / _\" \
/_\" \ / _\" \ / _\" \ / _\" \   version 3.5.0
  __/

Using Scala version 2.12.18, OpenJDK 64-Bit Server VM, 17.0.10
Branch HEAD
Compiled by user ubuntu on 2023-09-09T01:53:20Z
Revision ce5ddad990373636e94071e7cef2f31021add07b
Url https://github.com/apache/spark
Type --help for more information.
I have no name!@spark-master:/opt/bitnami/spark$
```

7) Выведите web-страницу с информацией о Spark и его текущей загрузке

8) Сконфигурируйте и подготовьте к работе Spark SQL