

Steps

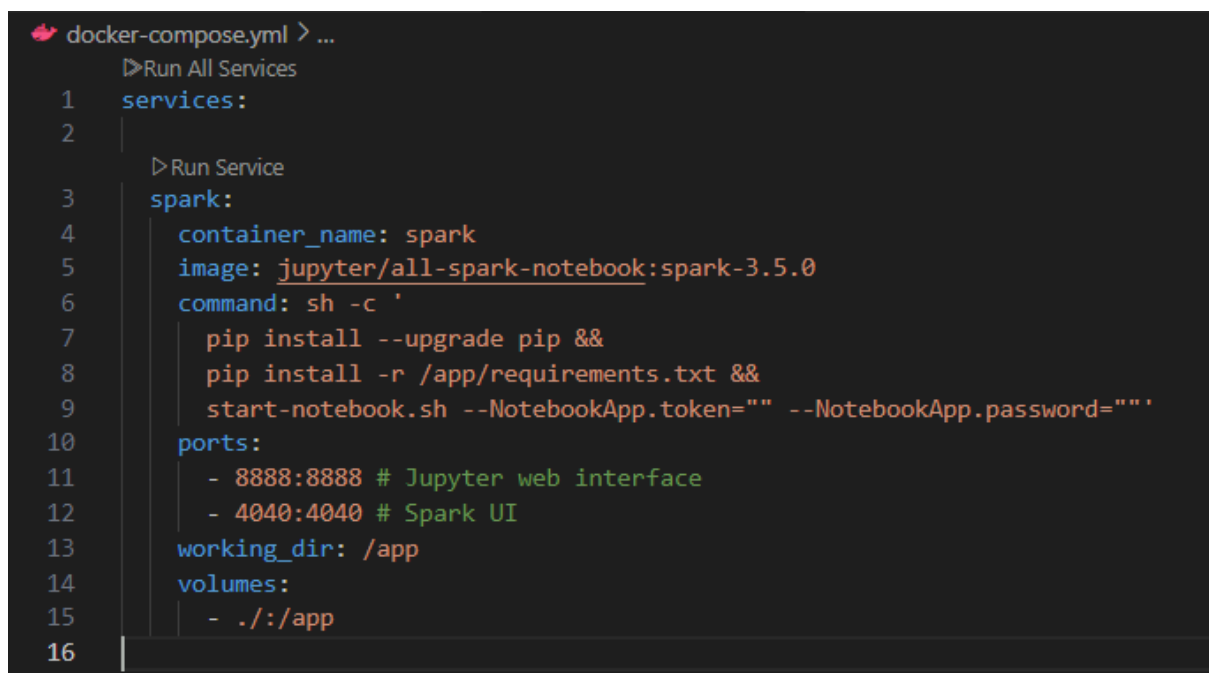
- Start docker containers

```
docker compose up --build
```

- Run clusterization algorithm

```
docker exec spark python src/clusterize.py
```

Docker конфигурация



```
docker-compose.yml > ...
  ▶ Run All Services
1  services:
2
3    ▶ Run Service
4    spark:
5      container_name: spark
6      image: jupyter/all-spark-notebook:spark-3.5.0
7      command: sh -c '
8        pip install --upgrade pip &&
9        pip install -r /app/requirements.txt &&
10       start-notebook.sh --NotebookApp.token="" --NotebookApp.password=""
11     ports:
12       - 8888:8888 # Jupyter web interface
13       - 4040:4040 # Spark UI
14     working_dir: /app
15     volumes:
16       - ./:/app
```

Конфигурация спарк приложения

```
spark.ini
conf > spark.ini > ...
1 [spark]
2 spark.master = local[*]
3 spark.driver.memory = 2g
4 spark.executor.memory = 1g
5 spark.executor.instances = 2
6 spark.executor.cores = 2
7 spark.dynamicAllocation.enabled = true
8 spark.dynamicAllocation.minExecutors = 1
9 spark.dynamicAllocation.maxExecutors = 5
10 spark.sql.execution.arrow.pyspark.enabled = true
11
12 [model]
13 seed = 42
14 k = 5
```

Состояние контейнеров в Docker Desktop

Containers [Give feedback](#)

View all your running containers and applications. [Learn more](#)

Container CPU usage 2.00% / 300% (3 CPUs available)

Container memory usage 5.55GB / 7.37GB

Show charts

Q Search

Only show running containers

<input type="checkbox"/>	Name	Container ID	Image	Port(s)	CPU (%)	Last started	Actions
<input type="checkbox"/>	lab-5	-	-	-	29.06%	1 hour ago	<div><div></div><div></div><div></div></div>
<input type="checkbox"/>	spark	3ea6dffa6841	jupyter/all-spark-notebook:spark-3	4040:4040 Show all ports (2)	29.06%	1 hour ago	<div><div></div><div></div><div></div></div>

История вычислений в Spark UI

Spark Jobs ^(?)

User: jovyann
Total Uptime: 48 min
Scheduling Mode: FIFO
Completed Jobs: 56

▶ Event Timeline

Completed Jobs (56)

Page: 1 1 Pages, jump to 1 . Show 100 items in a page. Go

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
55	save at NativeMethodAccessorImpl.java:0 save at NativeMethodAccessorImpl.java:0	2025/05/03 09:08:30	33 s	1/1 (2 skipped)	3/3 (164 skipped)
54	save at NativeMethodAccessorImpl.java:0 save at NativeMethodAccessorImpl.java:0	2025/05/03 09:05:03	3.4 min	1/1	82/82
53	save at NativeMethodAccessorImpl.java:0 save at NativeMethodAccessorImpl.java:0	2025/05/03 09:05:02	2.2 min	1/1	82/82
52	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2025/05/03 09:01:51	5 s	1/1 (2 skipped)	1/1 (164 skipped)
51	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2025/05/03 08:59:46	2.0 min	1/1	82/82
50	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2025/05/03 08:59:45	46 s	1/1	82/82
49	collect at ClusteringMetrics.scala:102 collect at ClusteringMetrics.scala:102	2025/05/03 08:58:38	0.2 s	1/1 (1 skipped)	1/1 (82 skipped)
48	collect at ClusteringMetrics.scala:102 collect at ClusteringMetrics.scala:102	2025/05/03 08:57:26	1.1 min	1/1	82/82
47	collectAsMap at ClusteringMetrics.scala:332 collectAsMap at ClusteringMetrics.scala:332	2025/05/03 08:55:54	1.5 min	2/2	164/164
46	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2025/05/03 08:55:49	2 s	1/1	1/1
45	collect at ClusteringSummary.scala:49 collect at ClusteringSummary.scala:49	2025/05/03 08:55:25	0.3 s	1/1 (1 skipped)	1/1 (82 skipped)
44	collect at ClusteringSummary.scala:49 collect at ClusteringSummary.scala:49	2025/05/03 08:53:09	2.2 min	1/1	82/82
43	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:59	5 s	2/2	164/164
42	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:57	2 s	2/2	164/164
41	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:55	2 s	2/2	164/164
40	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:52	3 s	2/2	164/164
39	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:50	2 s	2/2	164/164
38	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:46	3 s	2/2	164/164
37	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:43	3 s	2/2	164/164
36	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:39	3 s	2/2	164/164
35	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:36	3 s	2/2	164/164
34	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:34	2 s	2/2	164/164
33	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:32	1 s	2/2	164/164
32	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:29	3 s	2/2	164/164
31	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:27	2 s	2/2	164/164
30	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:23	3 s	2/2	164/164
29	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:20	4 s	2/2	164/164
28	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:13	7 s	2/2	164/164
27	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:09	4 s	2/2	164/164
26	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:52:04	4 s	2/2	164/164
25	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:51:48	16 s	2/2	164/164
24	collectAsMap at KMeans.scala:331 collectAsMap at KMeans.scala:331	2025/05/03 08:51:24	23 s	2/2	164/164
23	countByValue at KMeans.scala:448 countByValue at KMeans.scala:448	2025/05/03 08:50:57	11 s	2/2	164/164
22	collect at KMeans.scala:425 collect at KMeans.scala:425	2025/05/03 08:50:53	4 s	1/1	82/82