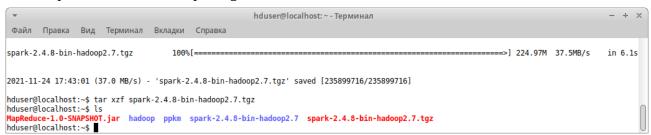
docker start -i gbhdp

wget https://apache-mirror.rbc.ru/pub/apache/spark/spark-2.4.8/spark-2.4.8-bin-hadoop2.7.tgz

tar xzf spark-2.4.8-bin-hadoop2.7.tgz



mv spark-2.4.8-bin-hadoop2.7 spark

```
hduser@localhost:~-Терминал — + ×

Файл Правка Вид Терминал Вкладки Справка
hduser@localhost:~$ tar xzf spark-2.4.8-bin-hadoop2.7.tgz
hduser@localhost:~$ ls

MapReduce-1.0-SNAPSHOT.jar hadoop ppkm spark-2.4.8-bin-hadoop2.7 spark-2.4.8-bin-hadoop2.7.tgz
hduser@localhost:~$ rm spark-2.4.8-bin-hadoop2.7.tgz
hduser@localhost:~$ ls

MapReduce-1.0-SNAPSHOT.jar hadoop ppkm spark-2.4.8-bin-hadoop2.7
hduser@localhost:~$ mv spark-2.4.8-bin-hadoop2.7
hduser@localhost:~$ ls

MapReduce-1.0-SNAPSHOT.jar hadoop ppkm spark
hduser@localhost:~$ Is

MapReduce-1.0-SNAPSHOT.jar hadoop ppkm spark
hduser@localhost:~$ Is
```

Редактирование в vim .bashrc:

export SPARK_HOME=/home/hduser/spark export PATH=\$PATH:\$SPARK_HOME/bin export HADOOP_CONF_DIR=\$HADOOP_HOME/etc/hadoop

echo \$PATH

```
терминал Вкладки Справка

hduser@localhost:~-/spark$ cd ..

hduser@localhost:~/spark$ cd ..

hduser@localhost:~$ export SPARK_HOME=/home/hduser/spark

hduser@localhost:~$ export PATH=$PATH:$SPARK_HOME/bin

hduser@localhost:~$ export HAD00P_CONF_DIR=$HAD00P_HOME/etc/hadoop

hduser@localhost:~$ vim .bashrc

hduser@localhost:~$ vim .bashrc

hduser@localhost:~$ vim .bashrc

hduser@localhost:~$ park

/usr/local/sbin:/usr/local/bin:/usr/sbin:/sbin:/home/hduser/hadoop/s

bin:/home/hduser/.local/bin:/home/hduser/spark/bin

hduser@localhost:~$ ■
```

Запускаем приложение:

spark-submit

```
hduser@localhost: ~ - Терминал
                                                                                                       - + ×
 Файл
       Правка Вид Терминал Вкладки Справка
hduser@localhost:~$ spark-submit
Usage: spark-submit [options] <app jar | python file | R file> [app arguments]
Usage: spark-submit --kill [submission ID] --master [spark://...]
Usage: spark-submit --status [submission ID] --master [spark://...]
Usage: spark-submit run-example [options] example-class [example args]
Options:
  --master MASTER_URL
                               spark://host:port, mesos://host:port, yarn,
                               k8s://https://host:port, or local (Default: local[*]).
                              Whether to launch the driver program locally ("client") or
  --deploy-mode DEPLOY_MODE
                               on one of the worker machines inside the cluster ("cluster")
                               (Default: client).
  --class CLASS_NAME
                               Your application's main class (for Java / Scala apps).
  --name NAME
                               A name of your application.
  --jars JARS
                               Comma-separated list of jars to include on the driver
                               and executor classpaths.
                               Comma-separated list of maven coordinates of jars to include
  --packages
                               on the driver and executor classpaths. Will search the local
                               maven repo, then maven central and any additional remote
                               repositories given by --repositories. The format for the
                               coordinates should be groupId:artifactId:version.
  --exclude-packages
                               Comma-separated list of groupId:artifactId, to exclude while
                               resolving the dependencies provided in --packages to avoid
                               dependency conflicts.
  --repositories
                               Comma-separated list of additional remote repositories to
                               search for the maven coordinates given with --packages.
  --py-files PY FILES
                               Comma-separated list of .zip, .egg, or .py files to place
                               on the PYTHONPATH for Python apps.
  --files FILES
                               Comma-separated list of files to be placed in the working
                               directory of each executor. File paths of these files
                               in executors can be accessed via SparkFiles.get(fileName).
  --conf, -c PROP=VALUE
                               Arbitrary Spark configuration property.
                               Path to a file from which to load extra properties. If not
  --properties-file FILE
                               specified, this will look for conf/spark-defaults.conf.
  --driver-memory MEM
                              Memory for driver (e.g. 1000M, 2G) (Default: 1024M).
  --driver-java-options
--driver-library-path
                               Extra Java options to pass to the driver.
                               Extra library path entries to pass to the driver.
  --driver-class-path
                               Extra class path entries to pass to the driver. Note that
                               jars added with --jars are automatically included in the
                               classpath.
                              Memory per executor (e.g. 1000M, 2G) (Default: 1G).
  --executor-memory MEM
                               User to impersonate when submitting the application.
  --proxv-user NAME
                               This argument does not work with --principal / --keytab.
                               Show this help message and exit.
  --help, -h
  --verbose, -v
                               Print additional debug output.
  --version,
                               Print the version of current Spark.
 Cluster deploy mode only:
  --driver-cores NUM
                               Number of cores used by the driver, only in cluster mode
                               (Default: 1).
 Spark standalone or Mesos with cluster deploy mode only:
                              If given, restarts the driver on failure.
  --supervise
  --kill SUBMISSION ID
                              If given, kills the driver specified.
  --status SUBMISSION ID
                              If given, requests the status of the driver specified.
 Spark standalone and Mesos only:
  --total-executor-cores NUM Total cores for all executors.
 Spark standalone and YARN only:
  --executor-cores NUM
                               Number of cores per executor. (Default: 1 in YARN mode,
                               or all available cores on the worker in standalone mode)
 YARN-only:
  --queue QUEUE NAME
                               The YARN queue to submit to (Default: "default").
  --num-executors NUM
                               Number of executors to launch (Default: 2).
                               If dynamic allocation is enabled, the initial number of
                               executors will be at least NUM.
  --archives ARCHIVES
                               Comma separated list of archives to be extracted into the
                               working directory of each executor.
  --principal PRINCIPAL
                               Principal to be used to login to KDC, while running on
                               secure HDFS.
                               The full path to the file that contains the keytab for the
  --keytab KEYTAB
                               principal specified above. This keytab will be copied to
                               the node running the Application Master via the Secure
                               Distributed Cache, for renewing the login tickets and the
                               delegation tokens periodically.
hduser@localhost:~$
```

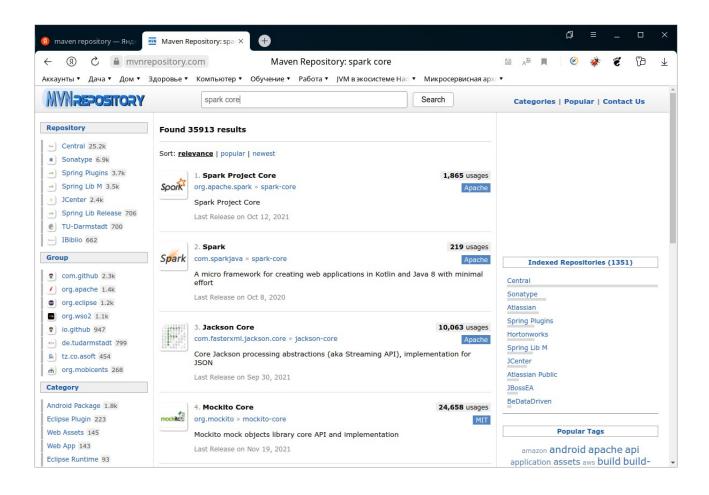
ls -la

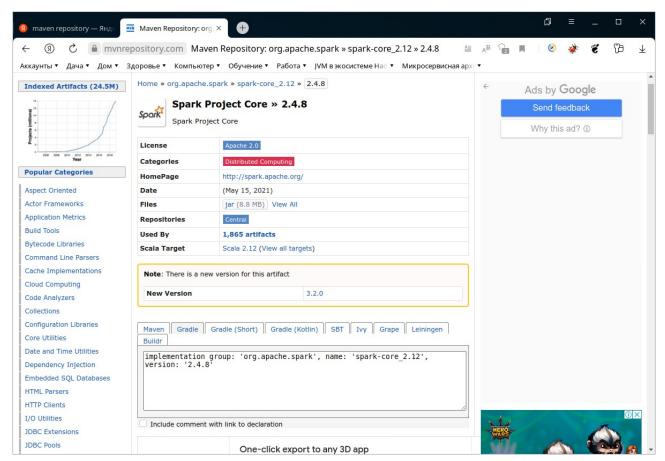
ls spark/

```
майл Правка Вид Терминал Вкладки Справка
drwxr-xr-x 13 hduser hduser 4096 May 8 2021 spark
hduser@localhost:~$ ls spark/
LICENSE R RELEASE conf examples kubernetes python yarn
NOTICE README.md bin data jars licenses sbin
hduser@localhost:~$ ■
```

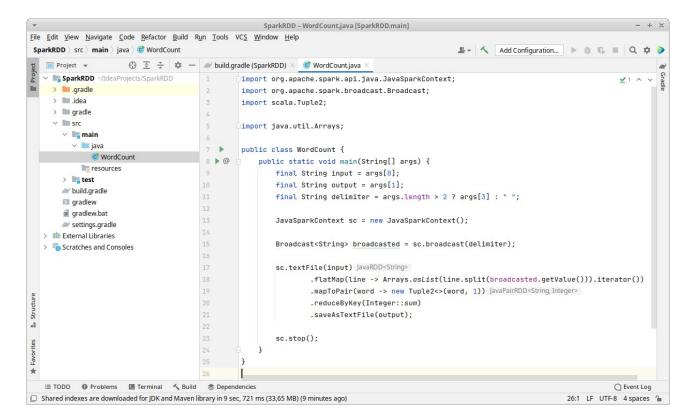
ls spark/jars/

hduser@localhost: ~ - Терминал Файл Правка Вид Терминал Вкладки Справка hduser@localhost:~\$ ls spark/jars/ jcl-over-slf4j-1.7.16.jar JavaEWAH-0.3.2.jar RoaringBitmap-0.7.45.jar ST4-4.0.4.jar jdo-api-3.0.1.jar jersey-client-2.22.2.jar activation-1.1.1.jar aircompressor-0.10.jar antlr-2.7.7.jar antlr-runtime-3.4.jar jersey-common-2.22.2.jar jersey-container-servlet-2.22.2.jar jersey-container-servlet-core-2.22.2.jar jersey-guava-2.22.2.jar antlr4-runtime-4.7.jar aopalliance-1.0.jar jersey-media-jaxb-2.22.2.jar jersey-server-2.22.2.jar jetty-6.1.26.jar aopalliance-repackaged-2.4.0-b34.jar apache-log4j-extras-1.2.17.jar apacheds-i18n-2.0.0-M15.jar jetty-util-6.1.26.jar jliné-2.14.6.jar joda-time-2.9.3.jar apacheds-kerberos-codec-2.0.0-M15.jar api-asn1-api-1.0.0-M20.jar jodd-core-3.5.2.jar api-util-1.0.0-M20.jar arpack_combined_all-0.1.jar arrow-format-0.10.0.jar jpam-1.1.jar json4s-ast_2.11-3.5.3.jar json4s-core_2.11-3.5.3.jar json4s-jackson_2.11-3.5.3.jar json4s-scalap_2.11-3.5.3.jar arrow-memory-0.10.0.jar arrow-vector-0.10.0.jar automaton-1.11-8.jar jsp-api-2.1.jar avro-1.8.2.jar avro-ipc-1.8.2.jar jsr305-1.3.9. jar jta-1.1. jar avro-mapred-1.8.2-hadoop2.jar bonecp-0.8.0.RELEASE.jar breeze-macros_2.11-0.13.2.jar breeze_2.11-0.13.2.jar calcite-avatica-1.2.0-incubating.jar jtransforms-2.4.0.jar jul-to-slf4j-1.7.16.jar kryo-shaded-4.0.2.jar kubernetes-client-4.6.1.jar kubernetes-model-4.6.1.jar calcite-core-1.2.0-incubating.jar
calcite-linq4j-1.2.0-incubating.jar kubernetes-model-common-4.6.1.jar leveldbjni-all-1.8.jar chill-java-0.9.3.jar chill_2.11-0.9.3.jar libfb303-0.9.3.jar libthrift-0.9.3.jar commons-beanutils-1.9.4.jar log4j-1.2.17.jar commons-cli-1.2.jar logging-interceptor-3.12.0.jar commons-codec-1.10.jar commons-collections-3.2.2.jar lz4-java-1.4.0.jar machinist_2.11-0.6.1.jar commons-compiler-3.0.16.jar commons-compress-1.20.jar macro-compat_2.11-1.1.1.jar mesos-1.4.0-shaded-protobuf.jar commons-configuration-1.6.jar commons-crypto-1.0.0.jar commons-dbcp-1.4.jar metrics-core-3.1.5.jar metrics-graphite-3.1.5.jar metrics-json-3.1.5.jar metrics-jvm-3.1.5.jar minlog-1.3.0.jar commons-digester-1.8.jar commons-httpclient-3.1.jar commons-io-2.4.jar commons-lang-2.6.jar netty-3.9.9.Final.jar netty-all-4.1.47.Final.jar commons-lang3-3.5.jar commons-logging-1.1.3.jar objenesis-2.5.1.jar okhttp-3.12.0.jar commons-math3-3.4.1.jar okio-1.15.0.jar commons-net-3.1.jar opencsv-2.3.jar commons-pool-1.5.4.jar compress-lzf-1.0.3.jar orc-core-1.5.5-nohive.jar orc-mapreduce-1.5.5-nohive.jar core-1.1.2.jar curator-client-2.7.1.jar orc-shims-1.5.5.jar oro-2.0.8.jar curator-framework-2.7.1.jar curator-recipes-2.7.1.jar osgi-resource-locator-1.0.1.jar paranamer-2.8.jar parquet-column-1.10.1.jar datanucleus-api-jdo-3.2.6.jar datanucleus-core-3.2.10.jar parquet-common-1.10.1.jar datanucleus-rdbms-3.2.9.jar parquet-encoding-1.10.1.jar derby-10.12.1.1.jar eigenbase-properties-1.1.5.jar flatbuffers-1.2.0-3f79e055.jar parquet-format-2.4.0.jar parquet-hadoop-1.10.1.jar parquet-hadoop-bundle-1.6.0.jar parquet-jackson-1.10.1.jar generex-1.0.2.jar gson-2.2.4.jar guava-14.0.1.jar protobuf-java-2.5.0.jar py4j-0.10.7.jar guava-14.0.1.jar guice-3.0.jar guice-servlet-3.0.jar hadoop-annotations-2.7.3.jar hadoop-auth-2.7.3.jar pyrolite-4.13.jar scala-compiler-2.11.12.jar scala-library-2.11.12.jar scala-parser-combinators_2.11-1.1.0.jar hadoop-client-2.7.3.jar scala-reflect-2.11.12.jar hadoop-common-2.7.3.jar hadoop-hdfs-2.7.3.jar hadoop-mapreduce-client-app-2.7.3.jar hadoop-mapreduce-client-common-2.7.3.jar hadoop-mapreduce-client-core-2.7.3.jar hadoop-mapreduce-client-jobclient-2.7.3.jar scala-xml_2.11-1.0.5.jar shapeless_2.11-2.3.2.jar shims-0.7.45.jar slf4j-api-1.7.16.jar slf4j-log4j12-1.7.16.jar snakeyaml-1.15.jar snappy-0.2.jar snappy-java-1.1.8.2.jar hadoop-mapreduce-client-shuffle-2.7.3.jar hadoop-yarn-api-2.7.3.jar snappy-java-1.1.0.2.jar spark-catalyst_2.11-2.4.8.jar spark-graphx_2.11-2.4.8.jar spark-hive-thriftserver_2.11-2.4.8.jar hadoop-yarn-client-2.7.3.jar hadoop-yarn-common-2.7.3.jar hadoop-yarn-server-common-2.7.3.jar hadoop-yarn-server-web-proxy-2.7.3.jar hive-beeline-1.2.1.spark2.jar hive-cli-1.2.1.spark2.jar hive-exec-1.2.1.spark2.jar hive-jdbc-1.2.1.spark2.jar hive-metastore-1.2.1.spark2.jar spark-hive 2.11-2.4.8.jar spark-kubernetes 2.11-2.4.8.jar spark-kvstore 2.11-2.4.8.jar spark-launcher 2.11-2.4.8.jar spark-mesos 2.11-2.4.8.jar spark-mllib-local 2.11-2.4.8.jar hk2-api-2.4.0-b34.jar hk2-locator-2.4.0-b34.jar spark-mllib_2.11-2.4.8.jar hk2-utils-2.4.0-b34.jar spark-network-common 2.11-2.4.8.jar spark-network-shuffle 2.11-2.4.8.jar hppc-0.7.2.jar htrace-core-3.1.0-incubating.jar httpclient-4.5.6.jar spark-repl_2.11-2.4.8.jar spark-sketch_2.11-2.4.8.jar spark-sql_2.11-2.4.8.jar spark-streaming_2.11-2.4.8.jar spark-tags_2.11-2.4.8.jar spark-tags_2.11-2.4.8.jar spark-unsafe_2.11-2.4.8.jar spark-yarn_2.11-2.4.8.jar httpcore-4.4.10.jar ivy-2.4.0.jar lvy-2.4.0.jar jackson-annotations-2.6.7.jar jackson-core-2.6.7.jar jackson-core-asl-1.9.13.jar jackson-databind-2.6.7.3.jar jackson-dataformat-yaml-2.6.7.jar jackson-jaxrs-1.9.13.jar spire-macros_2.11-0.13.0.jar spire_2.11-0.13.0.jar Jackson-mapper-asl-1.9.13.jar jackson-module-jaxb-annotations-2.6.7.jar jackson-module-paranamer-2.7.9.jar jackson-module-scala 2.11-2.6.7.1.jar stax-api-1.0-2.jar stax-api-1.0.1.jar stream-2.7.0.jar stringtemplate-3.2.1.jar super-csv-2.2.0.jar jackson-xc-1.9.13.jar univocity-parsers-2.7.3.jar validation-api-1.1.0.Final.jar janino-3.0.16.jar javassist-3.18.1-GA.jar xbean-asm6-shaded-4.8.jar javax.annotation-api-1.2.jar javax.inject-1.jar xercesImpl-2.9.1.jar javax.inject-2.4.0-b34.jar javax.servlet-api-3.1.0.jar xmlenc-0.52.jar xz-1.5.jar javax.ws.rs-api-2.0.1.jar javolution-5.5.1.jar zjsonpatch-0.3.0.jar zookeeper-3.4.6.jar zstd-jni-1.4.4-3.jar jaxb-api-2.2.2.jai hduser@localhost:~\$

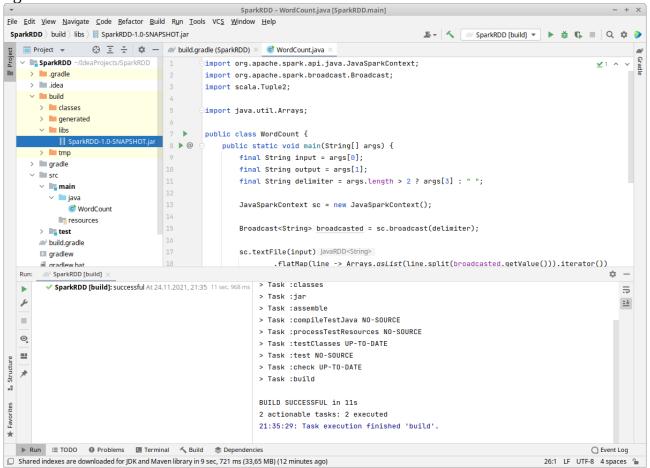




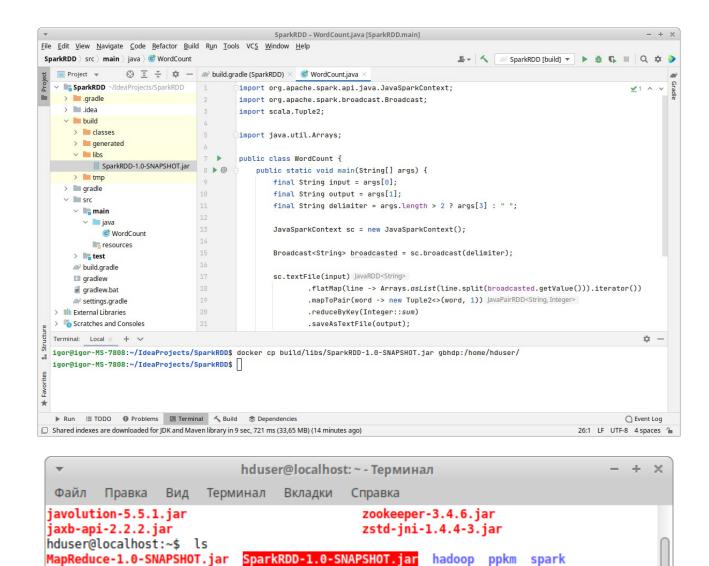
Создадим WordCount.java:



./gradlew build



docker cp build/libs/SparkRDD-1.0-SNAPSHOT.jar gbhdp:/home/hduser/



JAVA WORD COUNT: ЗАПУСК

Запускаем приложение:

hduser@localhost:~\$

```
hduser@localhost: ~ - Терминал
    Файл Правка Вид Терминал Вкладки Справка
 hduser@localhost:~$ spark-submit \
  > --class WordCount \
 > --master yarn \
      --deploy-mode cluster
 > SparkRDD-1.0-SNAPSHOT.jar \
  > /user/hduser/ppkm/ppkm_dataset.csv /user/hduser/ppkm-rdd-out
 21/11/24 18:41:08 WARN util.Utils: Your hostname. localhost resolves to a loopback address: 127.0.0.1: using 172.17.0.2 instead (on int
 21/11/24 18:41:08 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
 21/11/24 18:41:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes wher
  e applicable
 21/11/24 18:41:09 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032 21/11/24 18:41:10 INFO yarn.Client: Requesting a new application from cluster with 1 NodeManagers 21/11/24 18:41:10 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of the cluster
 (2048 MB per container)
21/11/24 18:41:10 INFO yarn.Client: Will allocate AM container, with 1408 MB memory including 384 MB overhead
21/11/24 18:41:10 INFO yarn.Client: Setting up container launch context for our AM
21/11/24 18:41:10 INFO yarn.Client: Setting up the launch environment for our AM container
21/11/24 18:41:10 INFO yarn.Client: Preparing resources for our AM container
21/11/24 18:41:10 INFO yarn.Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SP
 21/11/24 18:41:12 INFO yarn.Client: Uploading resource file:/tmp/spark-8b54a891-d3df-4ff8-b32f-caf25f54af5a/_spark_libs_2089933854880 916409.zip -> hdfs://localhost:9000/user/hduser/.sparkStaging/application_1637774987878_0001/_spark_libs_2089933854880916409.zip 21/11/24 18:41:13 INFO yarn.Client: Uploading resource file:/home/hduser/SparkRDD-1.0-SNAPSHOT.jar -> hdfs://localhost:9000/user/hduser/.sparkStaging/application_1637774987878_0001/SparkRDD-1.0-SNAPSHOT.jar 21/11/24 18:41:13 INFO yarn.Client: Uploading resource file:/tmp/spark-8b54a891-d3df-4ff8-b32f-caf25f54af5a/_spark_conf_2289161166840
  616666.zip -> hdfs://localhost:9000/user/hduser/.sparkStaging/application_1637774987878_0001/__spark_conf__.zip
 21/11/24 18:41:13 INFO spark.SecurityManager: Changing view acts to: hduser 21/11/24 18:41:13 INFO spark.SecurityManager: Changing modify acts to: hduser
 21/11/24 18:41:13 INFO spark.SecurityManager: Changing view acls groups to:
21/11/24 18:41:13 INFO spark.SecurityManager: Changing modify acls groups to:
21/11/24 18:41:13 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions:
 Set(hduser); groups with view permissions: Set(); users with modify permissions: Set(hduser); groups with modify permissions: Set() 21/11/24 18:41:14 INFO yarn.Client: Submitting application application 1637774987878 0001 to ResourceManager 21/11/24 18:41:15 INFO impl.YarnClientImpl: Submitted application application 1637774987878 0001 (state: ACCEPTED)
 21/11/24 18:41:16 INFO yarn.Client:
client token: N/A
                         diagnostics: AM container is launched, waiting for AM container to Register with RM
                         ApplicationMaster host: N/A
ApplicationMaster RPC port: -1
                         queue: default
                         start time: 1637779274948
final status: UNDEFINED
                          tracking URL: http://localhost:8088/proxy/application_1637774987878_0001/
user: hduser

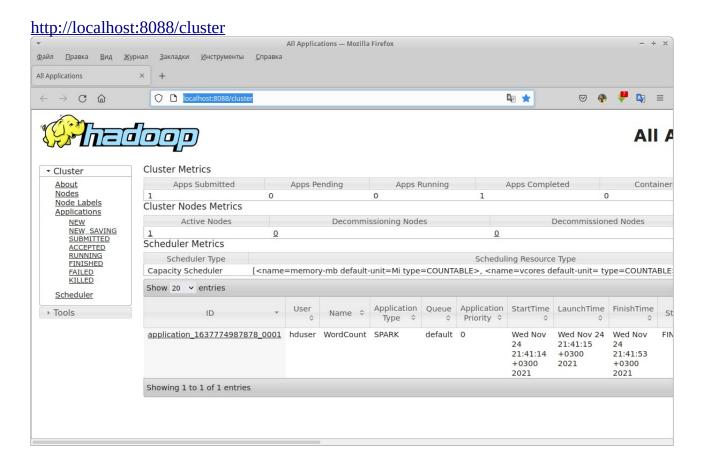
21/11/24 18:41:17 INFO yarn.Client: Application report for application_1637774987878_0001 (state: ACCEPTED)

21/11/24 18:41:18 INFO yarn.Client: Application report for application_1637774987878_0001 (state: ACCEPTED)

21/11/24 18:41:19 INFO yarn.Client: Application report for application_1637774987878_0001 (state: ACCEPTED)

21/11/24 18:41:20 INFO yarn.Client: Application report for application_1637774987878_0001 (state: ACCEPTED)

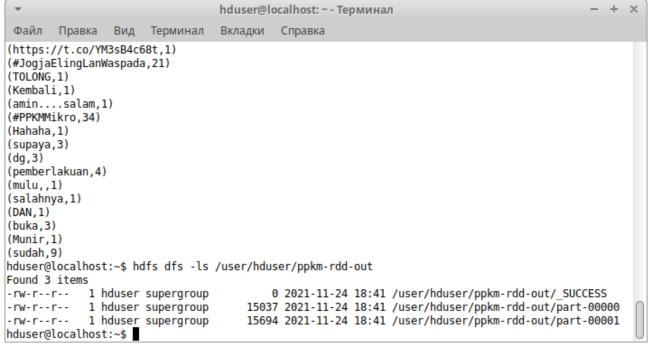
21/11/24 18:41:21 INFO yarn.Client: Application report for application_1637774987878_0001 (state: RUNNING)
 21/11/24 18:41:21 INFO yarn.Client:
client token: N/A
                         diagnostics: N/A
                         ApplicationMaster host: localhost
ApplicationMaster RPC port: 44291
                         queue: default
                         start time: 1637779274948 final status: UNDEFINED
                         tracking URL: http://localhost:8088/proxy/application_1637774987878_0001/
                         user: hduser
 21/11/24 18:41:22 INFO yarn.Client: Application report for application_1637774987878_0001 (state: RUNNING) 21/11/24 18:41:23 INFO yarn.Client: Application report for application_1637774987878_0001 (state: RUNNING)
 21/11/24 18:41:24 INFO yarn.Client: Application report for application 1637774987878 0001 (state: RUNNING) 21/11/24 18:41:25 INFO yarn.Client: Application report for application 1637774987878_0001 (state: RUNNING) 21/11/24 18:41:26 INFO yarn.Client: Application report for application_1637774987878_0001 (state: RUNNING)
 21/11/24 18:41:27 INFO yarn.Client: Application report for application 1637774987878 0001 (state: RUNNING) 21/11/24 18:41:28 INFO yarn.Client: Application report for application 1637774987878_0001 (state: RUNNING) 21/11/24 18:41:29 INFO yarn.Client: Application report for application_1637774987878_0001 (state: RUNNING)
21/11/24 18:41:30 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:31 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:32 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:33 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:33 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:35 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:35 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:37 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:37 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:38 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:39 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:40 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:41 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:41 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:43 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:43 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:44 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:45 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUNNING) 21/11/24 18:41:45 INFO yarn.Client: Application report for application 1637774987878 e0001 (state: RUN
 21/11/24 18:41:46 INFO yarn.Client: Application report for application 1637774987878 0001 (state: RUNNING) 21/11/24 18:41:47 INFO yarn.Client: Application report for application_1637774987878_0001 (state: RUNNING) 21/11/24 18:41:48 INFO yarn.Client: Application report for application_1637774987878_0001 (state: RUNNING)
 21/11/24 18:41:49 INFO yarn.Client: Application report for application 1637774987878 0001 (state: RUNNING) 21/11/24 18:41:50 INFO yarn.Client: Application report for application_1637774987878_0001 (state: RUNNING) 21/11/24 18:41:51 INFO yarn.Client: Application report for application_1637774987878_0001 (state: RUNNING)
 21/11/24 18:41:52 INFO yarn.Client: Application report for application_1637774987878_0001 (state: RUNNING) 21/11/24 18:41:53 INFO yarn.Client: Application report for application_1637774987878_0001 (state: RUNNING) 21/11/24 18:41:54 INFO yarn.Client: Application report for application_1637774987878_0001 (state: FINISHED)
 21/11/24 18:41:54 INFO yarn.Client:
client token: N/A
                         diagnostics: N/A
                         ApplicationMaster host: localhost
                         ApplicationMaster RPC port: 44291
                         queue: default
                         start time: 1637779274948
                         final status: SUCCEEDED
                         tracking URL: http://localhost:8088/proxy/application_1637774987878_0001/
                         user: hduser
  21/11/24 18:41:54 INFO util.ShutdownHookManager: Shutdown hook called
 21/11/24 18:41:54 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-d266697e-66c1-4021-beb8-ad7cf3938f6d 21/11/24 18:41:54 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-8b54a891-d3df-4ff8-b32f-caf25f54af5a
  hduser@localhost:~$
```



Смотрим результат:

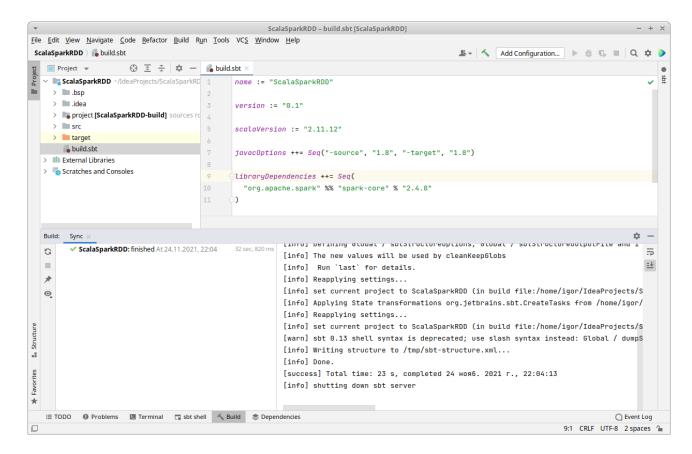
\$ hdfs dfs -ls /user/hduser/ppkm-rdd-out

\$ hdfs dfs -cat /user/hduser/ppkm-rdd-out/*
hduser@localho

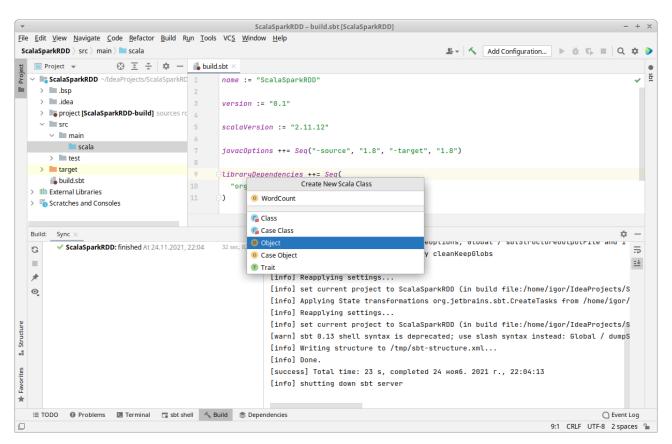


SCALA WORD COUNT: ΠΡΟΕΚΤ Β IDEA

Добавим в build.sbt зависимость на Apache Spark:



SCALA WORD COUNT: КОД



Создадим WordCount.scala:

```
ScalaSparkRDD - WordCount.scala [ScalaSparkRDD]
<u>F</u>ile <u>E</u>dit <u>V</u>iew <u>N</u>avigate <u>C</u>ode <u>R</u>efactor <u>B</u>uild <u>Run <u>T</u>ools VC<u>S</u> <u>W</u>indow <u>H</u>elp</u>
ScalaSparkRDD > src > main > scala > 0 WordCount.scala
                                                                                                   ■ Project ▼
                                                                                                                                                  sbt
   ✓ ■ ScalaSparkRDD ~/IdeaProjects/ScalaSparkRD 1
                                               import org.apache.spark.SparkContext
    > 🖿 .bsp
                                        3 ▶ object WordCount {
     > 📭 project [ScalaSparkRDD-build] sources ro 4
                                                def main(args: Array[String]): Unit = {
     ∨ 🗎 src
                                                  val input = args(0)
                                                   val output = args(1)
     ∨ 📄 scala
             WordCount
                                                  val sc = new SparkContext()
    > limitarget
                                                  sc.textFile(input)
       🔓 build.sbt
                                                     .flatMap(line => line.split( regex = "\\s"))
  > Illi External Libraries
                                                    .map(word => (word, 1))
    Scratches and Consoles
                                                    .reduceByKey(_ + _) // .reduceByKey((x, y) => x + y)
                                                    .sortBy { case (word, count) => count } // .sortBy(_._2) or .sortBy(pair => pair._2)
                                                    .map { case (word, count) => s"$word\t$count"}
... Structure
                                                    .saveAsTextFile(output)
                                                   sc.stop()
   C Event Log
                                                                                                                            20:2 LF UTF-8 2 spaces 🦫
```

SCALA WORD COUNT: СБОРКА И ДОСТАВКА

Соберем jar: \$ sbt package

```
ScalaSparkRDD - WordCount.scala [ScalaSparkRDD]
                                                                                                                                             - + ×
File Edit View Navigate Code Refactor Build Run Tools VCS Window Help
\textbf{ScalaSparkRDD} \ \rangle \ \ \text{target} \ \rangle \ \ \text{scala-2.11} \ \rangle \ \| \ \ \text{scalasparkrdd\_2.11-0.1.jar}
                                                                                                    🎩 マ 🔨 Add Configuration... ▶ 🇯 🖏 🔳 Q 🌣 🕨
                     ⊕ 🚊 😤 💠 — 🐔 build.sbt × 🧿 WordCount.scala >
                                                                                                                                          A1 ^ v
  ✓ ■ ScalaSparkRDD ~/IdeaProjects/ScalaSparkRD 1
                                              import org.apache.spark.SparkContext
   > lim .bsp
    > 🗎 .idea
                                             object WordCount {
    > 🃭 project [ScalaSparkRDD-build] sources rd 4 🕨
                                                def main(args: Array[String]): Unit = {
    ∨ 🗎 src
                                                  val input = args(0)

∨ IIII main

                                                  val output = args(1)
        ∨ ■ scala
             WordCount
                                                  val sc = new SparkContext()
       > 🗎 test

∨ I target

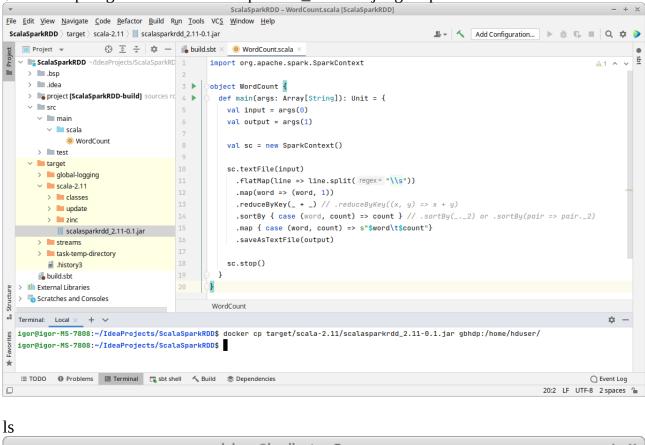
                                                  sc.textFile(input)
      > global-logging
                                                   .flatMap(line => line.split( regex = "\\s"))

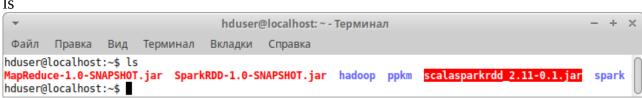
∨ I scala-2.11

                                                    .map(word => (word, 1))
        > III classes
                                                    .reduceByKey(_ + _) // .reduceByKey((x, y) => x + y)
         > update
                                                   .sortBy { case (word, count) => count } // .sortBy(_._2) or .sortBy(pair => pair._2)
         > inc
                                                    .map { case (word, count) => s"$word\t$count"}
         scalasparkrdd_2.11-0.1.jar
                                                   .saveAsTextFile(output)
      > iii streams
      > task-temp-directory
         .history3
       build.sbt
  > Illi External Libraries
   Scratches and Consoles
  sbt shell ScalaSparkRDD
  [info] set current project to ScalaSparkRDD (in build file:/home/igor/IdeaProjects/ScalaSparkRDD/)
   [info] Defining Global / ideaPort
     [info] The new value will be used by Compile / compile, Test / compile
     [info] Reapplying settings...
  🔛 [info] set current project to ScalaSparkRDD/)
  ⇒ [IJ]package
  [info] compiling 1 Scala source to /home/igor/IdeaProjects/ScalaSparkRDD/target/scala-2.11/classes ...
     [info] Non-compiled module 'compiler-bridge 2.11' for Scala 2.11.12. Compiling...
    [info] Compilation completed in 11.193s.
    [info] done compiling
...
    [success] Total time: 14 s, completed 24 Hos6. 2021 r., 22:09:59
    [IJ]
     >
   ○ Event Log
                                                                                                                              1:1 LF UTF-8 2 spaces 🦫
```

Перенесем его на кластер:

\$ docker cp target/scala-2.11/scalasparkrdd_2.11-0.1.jar gbhdp:/home/hduser/





spark-submit --class WordCount --master yarn --deploy-mode cluster scalasparkrdd_2.11-0.1.jar /user/hduser/ppkm /user/hduser/scala-ppkm-rdd-scala

```
hduser@localhost: ~ - Терминал
    Файл Правка Вид Терминал Вкладки Справка
  hduser@localhost:~$ spark-submit --class WordCount --master yarn --deploy-mode cluster scalasparkrdd 2.11-0.1.jar /user/hd
  user/ppkm /user/hduser/scala-ppkm-rdd-scala
  21/11/24 19:21:12 WARN util.Utils: Your hostname, localhost resolves to a loopback address: 127.0.0.1; using 172.17.0.2 in
  stead (on interface eth0)
  21/11/24 19:21:12 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
  21/11/24 19:21:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
 classes where applicable
21/11/24 19:21:13 INFO yarn.Client: Requesting a new application from cluster with 1 NodeManagers
21/11/24 19:21:13 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of
    the cluster (2048 MB per container)
 21/11/24 19:21:13 INFO yarn.Client: Will allocate AM container, with 1408 MB memory including 384 MB overhead 21/11/24 19:21:13 INFO yarn.Client: Setting up container launch context for our AM 21/11/24 19:21:13 INFO yarn.Client: Setting up the launch environment for our AM container 21/11/24 19:21:13 INFO yarn.Client: Preparing resources for our AM container 21/11/24 19:21:13 WARN yarn.Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libra
   ries under SPARK_HOME.
  21/11/24 19:21:15 INFO yarn.Client: Uploading resource file:/tmp/spark-f3f1e7e9-dc5d-4619-950b-318a8fddd3e7/_spark_libs_
1924761034981937149.zip -> hdfs://localhost:9000/user/hduser/.sparkStaging/application_1637774987878_0002/_spark_libs__19
  24761034981937149.zip
 21/11/24 19:21:16 INFO yarn.Client: Uploading resource file:/home/hduser/scalasparkrdd_2.11-0.1.jar -> hdfs://localhost:90 00/user/hduser/.sparkStaging/application_1637774987878_0002/scalasparkrdd_2.11-0.1.jar 21/11/24 19:21:16 INFO yarn.Client: Uploading resource file:/tmp/spark-f3f1e7e9-dc5d-4619-950b-318a8fddd3e7/__spark_conf_
  5472250497676915062.zip´-> hdfs://localhost:9000/user/hduser/.sparkStaging/application 1637774987878_0002/_spark_conf_.z
  21/11/24 19:21:16 INFO spark.SecurityManager: Changing view acls to: hduser
21/11/24 19:21:16 INFO spark.SecurityManager: Changing modify acls to: hduser
 22/11/24 19:21:16 INFO spark.SecurityManager: Changing view acts groups to:
21/11/24 19:21:16 INFO spark.SecurityManager: Changing modify acts groups to:
21/11/24 19:21:16 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acts disabled; users with view
    permissions: Set(hduser); groups with view permissions: Set(); users with modify permissions: Set(hduser); groups with m
 odify permissions: Set()
21/11/24 19:21:17 INFO yarn.Client: Submitting application application_1637774987878_0002 to ResourceManager
21/11/24 19:21:18 INFO impl.YarnClientImpl: Submitted application application_1637774987878_0002
21/11/24 19:21:18 INFO yarn.Client: Application report for application_1637774987878_0002 (state: ACCEPTED)
21/11/24 19:21:18 INFO yarn.Client:
                         client token: N/A
                         diagnostics: AM container is launched, waiting for AM container to Register with RM
                        ApplicationMaster host: N/A
ApplicationMaster RPC port: -1
                         queue: default
                         start time: 1637781677242
                         final status: UNDEFINED tracking URL: http://localhost:8088/proxy/application_1637774987878_0002/
                         user: hduser
 21/11/24 19:21:19 INFO yarn.Client: Application report for application_1637774987878_0002 (state: ACCEPTED) 21/11/24 19:21:20 INFO yarn.Client: Application report for application_1637774987878_0002 (state: ACCEPTED) 21/11/24 19:21:21 INFO yarn.Client: Application report for application_1637774987878_0002 (state: ACCEPTED)
 21/11/24 19:21:22 INFO yarn.Client: Application report for application_1637774987878_0002 (state: ACCEPTED) 21/11/24 19:21:23 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING) 21/11/24 19:21:23 INFO yarn.Client:
                         client token: N/A
                        diagnostics: N/A
ApplicationMaster host: localhost
                         ApplicationMaster RPC port: 32891
                         queue: default
                         start time: 1637781677242
                        final status: UNDEFINED tracking URL: http://localhost:8088/proxy/application_1637774987878_0002/
tracking URL: http://localhost:8088/proxy/application_1637774987878_0002/
user: hduser

21/11/24 19:21:24 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21:25 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21:26 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21:27 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21:28 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21:29 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21:30 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21:31 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21:33 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21:33 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21:33 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21:35 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21:35 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21:35 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

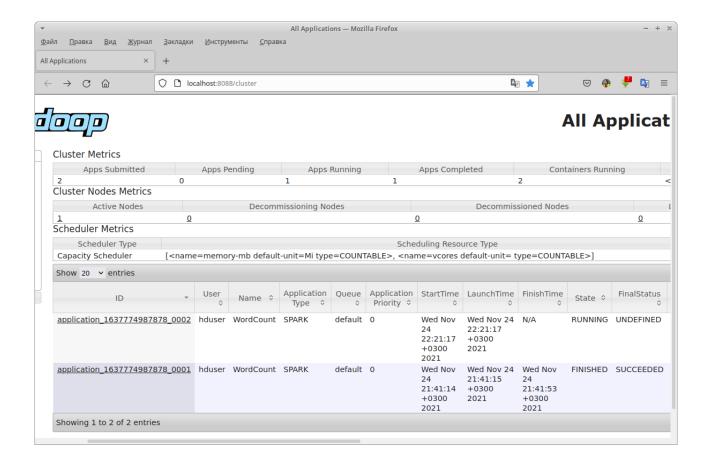
21/11/24 19:21:35 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21:37 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21:38 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21:39 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)

21/11/24 19:21
 21/11/24 19:21:49 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING) 21/11/24 19:21:40 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING) 21/11/24 19:21:42 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING) 21/11/24 19:21:43 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)
 21/11/24 19:21:44 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING) 21/11/24 19:21:45 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING) 21/11/24 19:21:46 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)
  21/11/24 19:21:47 INFO yarn.Client: Application report for application_1637774987878_0002 (state:
                                                                                                                                                                                                                                                          RUNNING)
 21/11/24 19:21:49 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING) 21/11/24 19:21:49 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING) 21/11/24 19:21:50 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING) 21/11/24 19:21:51 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING) 21/11/24 19:21:52 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING) 21/11/24 19:21:53 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING) 21/11/24 19:21:53 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING) 21/11/24 19:21:53 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING)
 21/11/24 19:21:54 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING) 21/11/24 19:21:55 INFO yarn.Client: Application report for application_1637774987878_0002 (state: RUNNING) 21/11/24 19:21:56 INFO yarn.Client: Application report for application_1637774987878_0002 (state: FINISHED) 21/11/24 19:21:56 INFO yarn.Client:
                         client token: N/A
                         diagnostics: N/A
                         ApplicationMaster host: localhost
                         ApplicationMaster RPC port: 32891
                         queue: default
start time: 1637781677242
                         final status: SUCCEEDED
                         tracking URL: http://localhost:8088/proxy/application_1637774987878_0002/
                         user: hduser
  21/11/24 19:21:56 INFO util.ShutdownHookManager: Shutdown hook called
 21/11/24 19:21:56 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-f3f1e7e9-dc5d-4619-950b-318a8fddd3e7 21/11/24 19:21:56 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-dac682d9-ddle-4986-993d-19f86b0378f8
  hduser@localhost:~$
```



Смотрим результат:

\$ hdfs dfs -ls /user/hduser/scala-ppkm-rdd-scala



\$ hdfs dfs -cat /user/hduser/scala-ppkm-rdd-scala/*

+	hduser@localhost: ~ - Терминал — + ×							
Файл					Cananya			
	Правка	Вид	Терминал	Вкладки	Справка			
22	11							
banyak	11							
hingga	11							
guna tidak	11 11							
100000000000000000000000000000000000000	11							
gara kembali								
aja	12							
rakyat	12							
kesehata		12						
dukung	12							
lagi	12							
Desa	12							
Masker	12							
sampai	12							
ini	13							
kebijaka		13						
kita	13							
bisa	13							
diperpar	njang	13						
protokol		14						
itu	15							
pak	15							
penyebar		15						
dari	16							
ppkm	16							
2021.	16							
orang	16							
mikro	18							
mulai	19							
atau	19	7.0						
#ppkmmik		19						
masyarak		20						
covid netral,F	20	21						
https://			C 21					
Pemberla		21	C 21					
#COVID-1		21						
Pengumun		21						
			21					
#JogjaElingLanWaspada 21 Mikro. 22								
Covid-19		23						
@humas_j		25						
#JogjaIs	stimewa	25						
Maret	26							
untuk	27							
yang	27							
ada	29							
Berbasis		30						
perpanja		32						
#PPKMMi		34						
Pembatasan		36 38						
	Masyarakat							
Kegiatar		39						
	41							
yg	55							
Mikro	56							
di	74							
PPKM	84							
dan	84	+. + =						
hduser@1	loca lnos	L:~\$						U