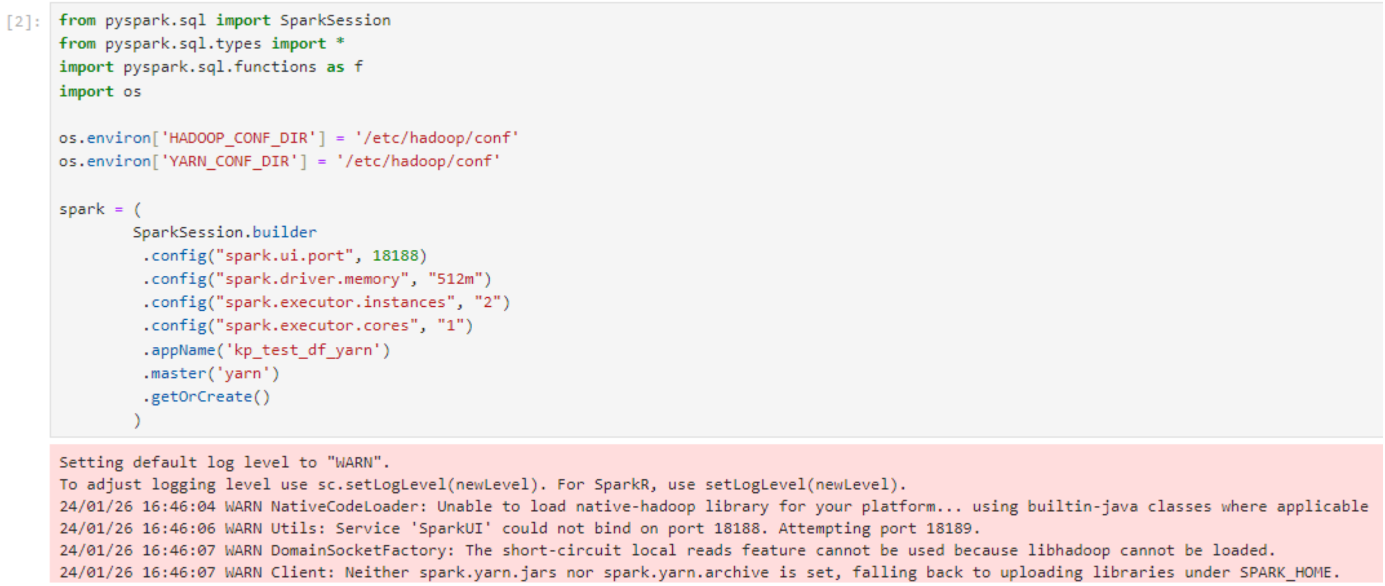
Задание по YARN

1. Подключитесь к Jupyter Notebook.

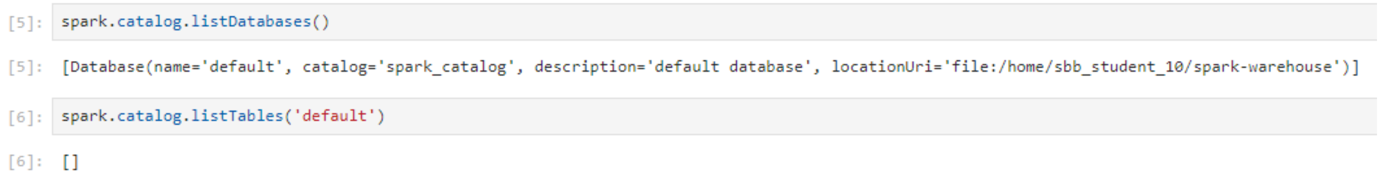
2. Создайте Spark сессию.



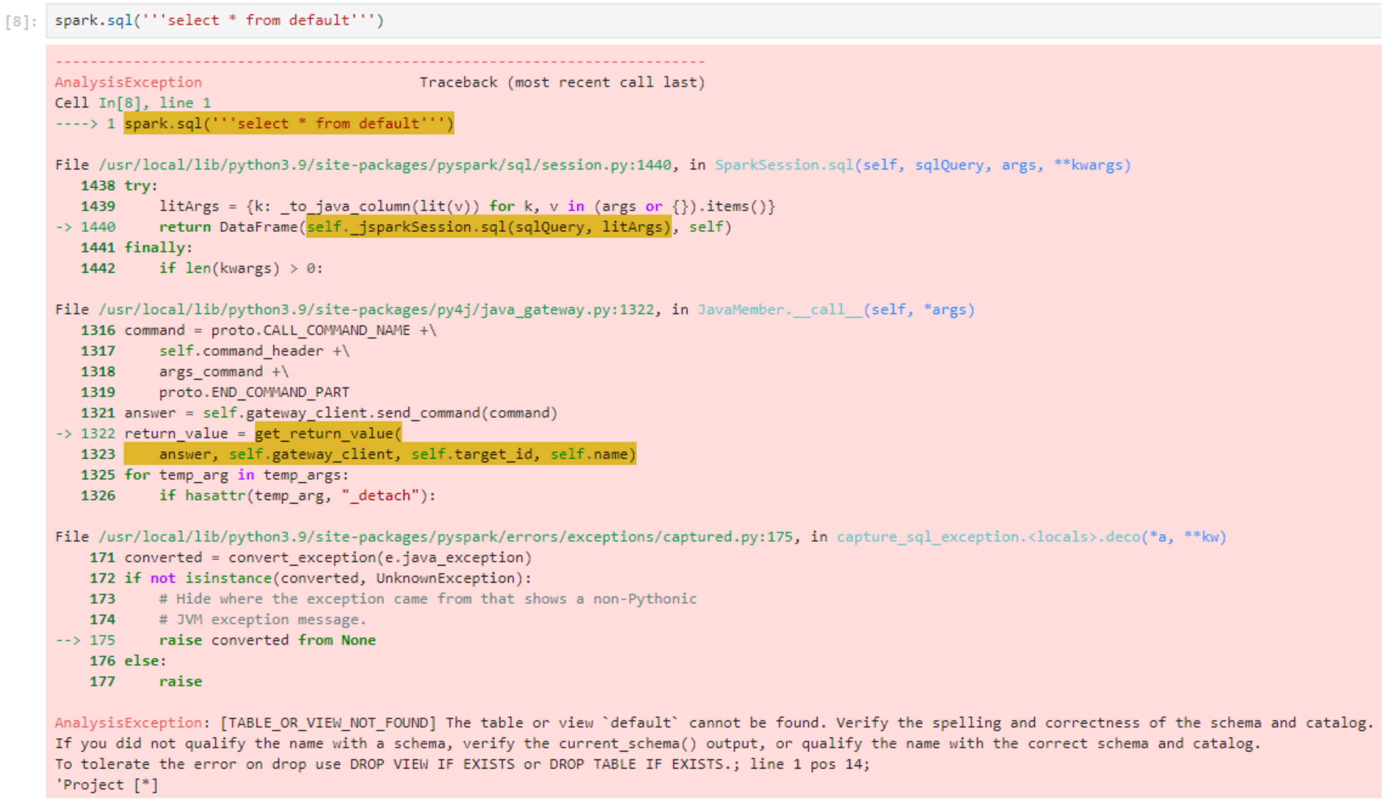
```
[2]: from pyspark.sql import SparkSession
     from pyspark.sql.types import *
     import pyspark.sql.functions as f
     import os

     os.environ['HADOOP_CONF_DIR'] = '/etc/hadoop/conf'
     os.environ['YARN_CONF_DIR'] = '/etc/hadoop/conf'

     spark = (
             SparkSession.builder
              .config("spark.ui.port", 18188)
              .config("spark.driver.memory", "512m")
              .config("spark.executor.instances", "2")
              .config("spark.executor.cores", "1")
              .appName('kp_test_df_yarn')
              .master('yarn')
              .getOrCreate()
             )
```

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/01/26 16:46:04 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/01/26 16:46:06 WARN Utils: Service 'SparkUI' could not bind on port 18188. Attempting port 18189.
24/01/26 16:46:07 WARN DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
24/01/26 16:46:07 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
```

3. При помощи spark.sql выполните пару запросов, например, выведите список баз данных, список таблиц в базе. Сделайте заведомо неправильный запрос (например, попробуйте выбрать данные из несуществующей таблицы).



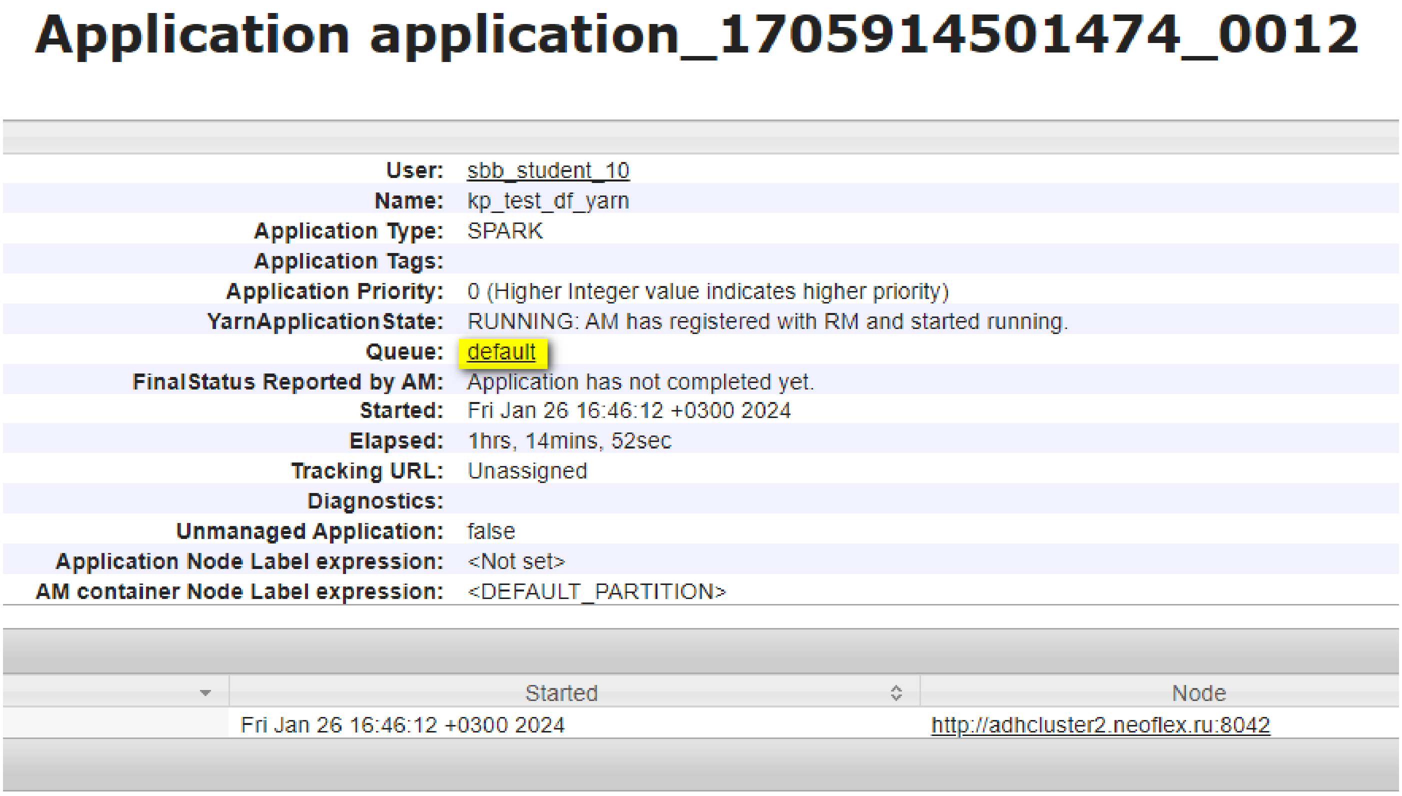
```
[5]: spark.catalog.listDatabases()

[5]: [Database(name='default', catalog='spark_catalog', description='default database', locationUri='file:/home/sbb_student_10/spark-warehouse')]

[6]: spark.catalog.listTables('default')

[6]: []
```



```
[8]: spark.sql('''select * from default''')
```

```
---------------------------------------------------------------------------
AnalysisException                         Traceback (most recent call last)
Cell In[8], line 1
----> 1 spark.sql('''select * from default''')

File /usr/local/lib/python3.9/site-packages/pyspark/sql/session.py:1440, in SparkSession.sql(self, sqlQuery, args, **kwargs)
   1438 try:
   1439     litArgs = {k: _to_java_column(lit(v)) for k, v in (args or {}).items()}
-> 1440     return DataFrame(self._jsparkSession.sql(sqlQuery, litArgs), self)
   1441 finally:
   1442     if len(kwargs) > 0:

File /usr/local/lib/python3.9/site-packages/py4j/java_gateway.py:1322, in JavaMember.__call__(self, *args)
   1316 command = proto.CALL_COMMAND_NAME +\
   1317     self.command_header +\
   1318     args_command +\
   1319     proto.END_COMMAND_PART
   1321 answer = self.gateway_client.send_command(command)
-> 1322 return_value = get_return_value(
   1323     answer, self.gateway_client, self.target_id, self.name)
   1325 for temp_arg in temp_args:
   1326     if hasattr(temp_arg, "_detach"):

File /usr/local/lib/python3.9/site-packages/pyspark/errors/exceptions/captured.py:175, in capture_sql_exception.<locals>.deco(*a, **kw)
    171 converted = convert_exception(e.java_exception)
    172 if not isinstance(converted, UnknownException):
    173     # Hide where the exception came from that shows a non-Pythonic
    174     # JVM exception message.
--> 175     raise converted from None
    176 else:
    177     raise

AnalysisException: [TABLE_OR_VIEW_NOT_FOUND] The table or view `default` cannot be found. Verify the spelling and correctness of the schema and catalog.
If you did not qualify the name with a schema, verify the current_schema() output, or qualify the name with the correct schema and catalog.
To tolerate the error on drop use DROP VIEW IF EXISTS or DROP TABLE IF EXISTS.; line 1 pos 14;
'Project [*]
```

4. Откройте интерфейс YARN, найдите свое задание Spark (по имени пользователя) и скопируйте ее номер (вида application_xxxxxxxxx_yyyyyyyy). Обратите внимание на название очереди, в которой запустилось задание. Сделайте скриншот и сохраните его в своей домашней папке.



Application application_1705914501474_0012

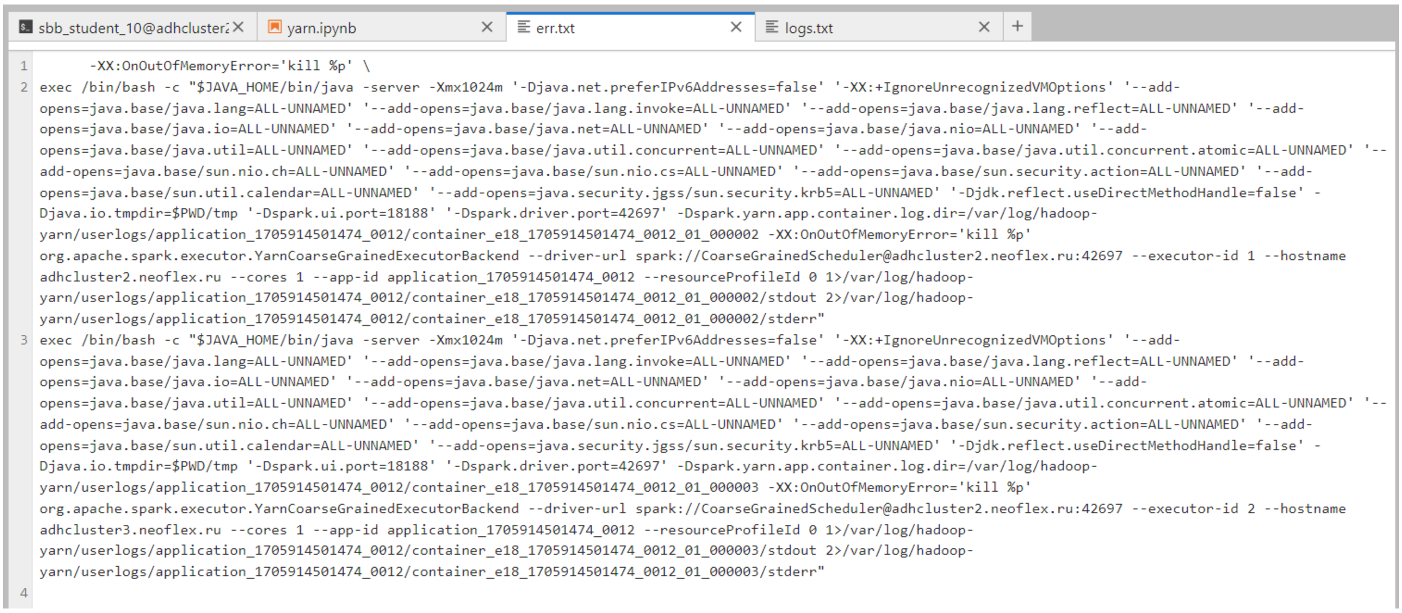
User:	sbb_student_10
Name:	kp_test_df_yarn
Application Type:	SPARK
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	RUNNING: AM has registered with RM and started running.
Queue:	default
FinalStatus Reported by AM:	Application has not completed yet.
Started:	Fri Jan 26 16:46:12 +0300 2024
Elapsed:	1hrs, 14mins, 52sec
Tracking URL:	Unassigned
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

	Started	Node
	Fri Jan 26 16:46:12 +0300 2024	http://adhcluster2.neoflex.ru:8042

5. В Jupyter Notebook откройте терминал.

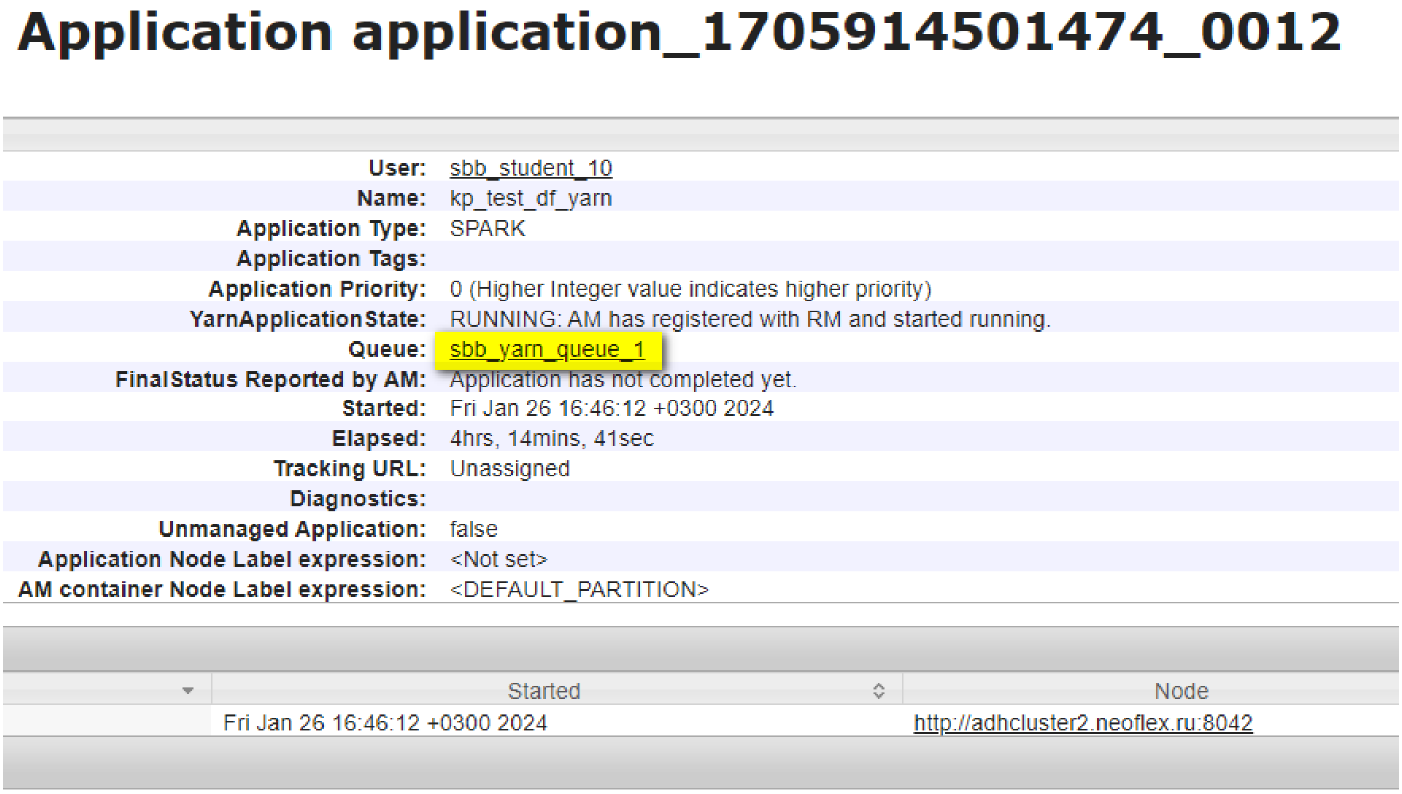
6. Выведите список всех заданий YARN, найдите среди них свое. (Подсказка, можно использовать команду linux grep для поиска)

7. Выведите лог своего задания только по строчкам, где есть слово Error. Сохраните результаты локально в файл в своей домашней директории.



```
      -XX:OnOutOfMemoryError='kill %p' \
exec /bin/bash -c "$JAVA_HOME/bin/java -server -Xmx1024m '-Djava.net.preferIPv6Addresses=false' '-XX:+IgnoreUnrecognizedVMOptions' '--add-
opens=java.base/java.lang=ALL-UNNAMED' '--add-opens=java.base/java.lang.invoke=ALL-UNNAMED' '--add-opens=java.base/java.lang.reflect=ALL-UNNAMED' '--add-
opens=java.base/java.io=ALL-UNNAMED' '--add-opens=java.base/java.net=ALL-UNNAMED' '--add-opens=java.base/java.nio=ALL-UNNAMED' '--add-
opens=java.base/java.util=ALL-UNNAMED' '--add-opens=java.base/java.util.concurrent=ALL-UNNAMED' '--add-opens=java.base/java.util.concurrent.atomic=ALL-UNNAMED' '--
add-opens=java.base/sun.nio.ch=ALL-UNNAMED' '--add-opens=java.base/sun.nio.cs=ALL-UNNAMED' '--add-opens=java.base/sun.security.action=ALL-UNNAMED' '--add-
opens=java.base/sun.util.calendar=ALL-UNNAMED' '--add-opens=java.security.jgss/sun.security.krb5=ALL-UNNAMED' '-Djdk.reflect.useDirectMethodHandle=false' -
Djava.io.tmpdir=$PWD/tmp '-Dspark.ui.port=18188' '-Dspark.driver.port=42697' -Dspark.yarn.app.container.log.dir=/var/log/hadoop-
yarn/userlogs/application_1705914501474_0012/container_e18_1705914501474_0012_01_000002 -XX:OnOutOfMemoryError='kill %p'
org.apache.spark.executor.YarnCoarseGrainedExecutorBackend --driver-url spark://CoarseGrainedScheduler@adhcluster2.neoflex.ru:42697 --executor-id 1 --hostname
adhcluster2.neoflex.ru --cores 1 --app-id application_1705914501474_0012 --resourceProfileId 0 1>/var/log/hadoop-
yarn/userlogs/application_1705914501474_0012/container_e18_1705914501474_0012_01_000002/stdout 2>/var/log/hadoop-
yarn/userlogs/application_1705914501474_0012/container_e18_1705914501474_0012_01_000002/stderr"
exec /bin/bash -c "$JAVA_HOME/bin/java -server -Xmx1024m '-Djava.net.preferIPv6Addresses=false' '-XX:+IgnoreUnrecognizedVMOptions' '--add-
opens=java.base/java.lang=ALL-UNNAMED' '--add-opens=java.base/java.lang.invoke=ALL-UNNAMED' '--add-opens=java.base/java.lang.reflect=ALL-UNNAMED' '--add-
opens=java.base/java.io=ALL-UNNAMED' '--add-opens=java.base/java.net=ALL-UNNAMED' '--add-opens=java.base/java.nio=ALL-UNNAMED' '--add-
opens=java.base/java.util=ALL-UNNAMED' '--add-opens=java.base/java.util.concurrent=ALL-UNNAMED' '--add-opens=java.base/java.util.concurrent.atomic=ALL-UNNAMED' '--
add-opens=java.base/sun.nio.ch=ALL-UNNAMED' '--add-opens=java.base/sun.nio.cs=ALL-UNNAMED' '--add-opens=java.base/sun.security.action=ALL-UNNAMED' '--add-
opens=java.base/sun.util.calendar=ALL-UNNAMED' '--add-opens=java.security.jgss/sun.security.krb5=ALL-UNNAMED' '-Djdk.reflect.useDirectMethodHandle=false' -
Djava.io.tmpdir=$PWD/tmp '-Dspark.ui.port=18188' '-Dspark.driver.port=42697' -Dspark.yarn.app.container.log.dir=/var/log/hadoop-
yarn/userlogs/application_1705914501474_0012/container_e18_1705914501474_0012_01_000003 -XX:OnOutOfMemoryError='kill %p'
org.apache.spark.executor.YarnCoarseGrainedExecutorBackend --driver-url spark://CoarseGrainedScheduler@adhcluster2.neoflex.ru:42697 --executor-id 2 --hostname
adhcluster3.neoflex.ru --cores 1 --app-id application_1705914501474_0012 --resourceProfileId 0 1>/var/log/hadoop-
yarn/userlogs/application_1705914501474_0012/container_e18_1705914501474_0012_01_000003/stdout 2>/var/log/hadoop-
yarn/userlogs/application_1705914501474_0012/container_e18_1705914501474_0012_01_000003/stderr"

```

8. Переместите свое задание в очередь «». Сделайте скриншот и сохраните его в домашней папке.

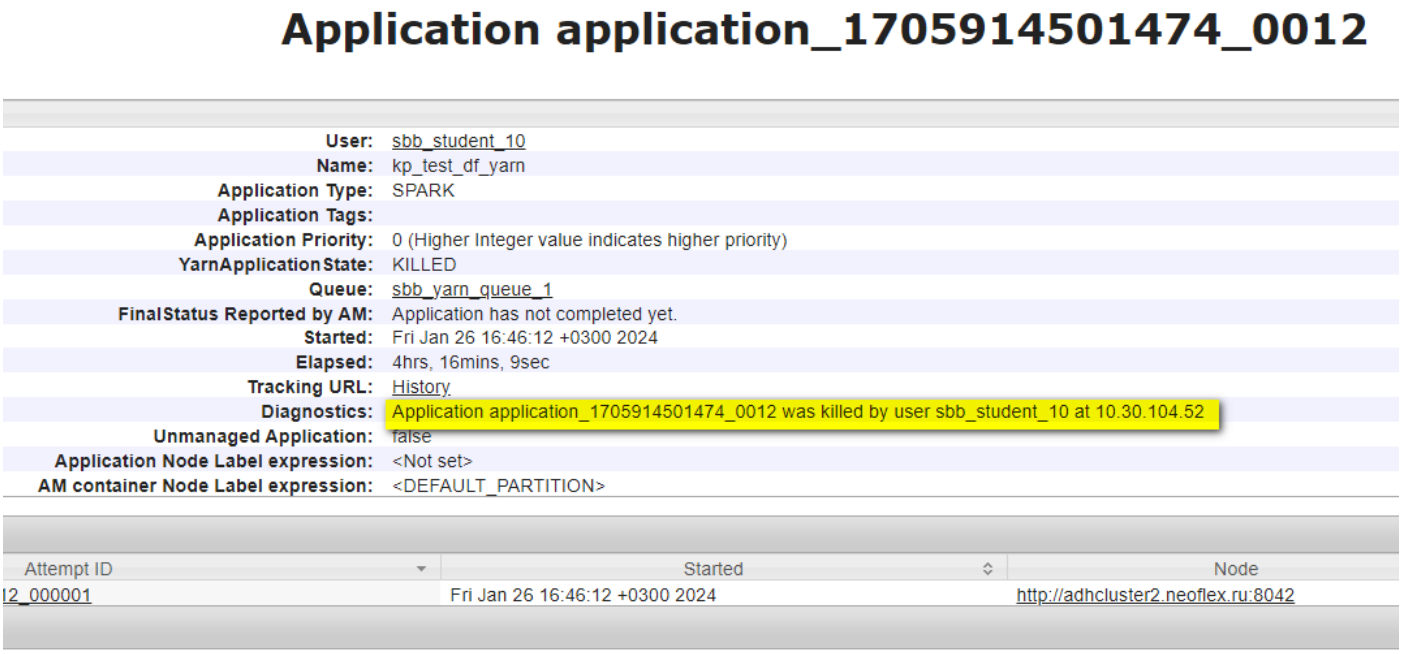


Application application_1705914501474_0012

User:	sbb_student_10
Name:	kp_test_df_yarn
Application Type:	SPARK
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	RUNNING: AM has registered with RM and started running.
Queue:	sbb_yarn_queue_1
FinalStatus Reported by AM:	Application has not completed yet.
Started:	Fri Jan 26 16:46:12 +0300 2024
Elapsed:	4hrs, 14mins, 41sec
Tracking URL:	Unassigned
Diagnostics:	
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

	Started	Node
	Fri Jan 26 16:46:12 +0300 2024	http://adhcluster2.neoflex.ru:8042

9. Прервите работу задания, сделайте скриншот и сохраните в домашней папке.



Application application_1705914501474_0012

User:	sbb_student_10
Name:	kp_test_df_yarn
Application Type:	SPARK
Application Tags:	
Application Priority:	0 (Higher Integer value indicates higher priority)
YarnApplicationState:	KILLED
Queue:	sbb_yarn_queue_1
FinalStatus Reported by AM:	Application has not completed yet.
Started:	Fri Jan 26 16:46:12 +0300 2024
Elapsed:	4hrs, 16mins, 9sec
Tracking URL:	History
Diagnostics:	Application application_1705914501474_0012 was killed by user sbb_student_10 at 10.30.104.52
Unmanaged Application:	false
Application Node Label expression:	<Not set>
AM container Node Label expression:	<DEFAULT_PARTITION>

Attempt ID	Started	Node
12_000001	Fri Jan 26 16:46:12 +0300 2024	http://adhcluster2.neoflex.ru:8042