# Hadoop-with-Oozie

≡ Руководство по выполнению тестового задания: Запустить и получить успешное завершение action в планировщике Apache Oozie

# Настройка виртуальной машины

- Скачиваем и устанавливаем ***VirtualBox***
- Скачиваем образ ***Ubuntu Server 22.04.2*** (.iso файл)
- Создаем виртуальную машину(далее ВМ) с характеристиками:
  - ОС - ***Ubuntu Server 22.04.2***
  - Процессор - ***2 ядра***
  - Оперативная память - ***4096 МБ***
  - Размер диска - ***25 ГБ***
  - Имя пользователя - ***hduser***
  - Выбираем ***Install OpenSSH server***
  - В настройках в разделе *Сеть* выбираем *Тип подключения* - ***Сетевой мост*** (для получения IP адреса из локальной сети)
- Вводим логин и пароль нашего пользователя
- Узнаем IP адрес нашей ВМ:

```
ip a
```

- Переходим в ***WSL***
- Генерируем пару ключей:

```
ssh-keygen
```

- Копируем публичный и приватный ключ в ВМ (копирование приватного ключа необходимо для правильной работы hadoop):

```
ssh-copy-id hduser@192.168.0.7
scp id_rsa hduser@192.168.0.7:~/.ssh
```

- Подключаемся к ВМ:

```
ssh hduser@192.168.0.7
```

- Проверяем можем ли подключится к локальному хосту по ***ssh***:

```
ssh localhost
```

- Обновляем пакеты:

```
sudo apt update
sudo apt upgrade
```

# Настройка Hadoop-2.6.0 Single Node Cluster

- В официальной документации на сайте ***Apache*** написано, что необходимо установить ***ssh*** и ***pdsh***. ***Ssh*** было установлено при создании ВМ. Устанавливаем ***pdsh***:

```
sudo apt install pdsh
```

- Также требуется ***java 8*** и ***maven***, устанавливаем:

```
sudo apt install openjdk-8-jre-headless openjdk-8-jdk
sudo apt install maven
```

- Скачиваем архив ***hadoop-2.6.0*** в домашнюю директорию пользователя:

```
wget https://archive.apache.org/dist/hadoop/core/hadoop-2.6.0/hadoop-
2.6.0.tar.gz
```

- Разархивируем скачанный файл:

```
sudo tar -xzvf hadoop-2.6.0.tar.gz
```

- Переименовываем директорию с разархивированным содержимым для удобства:

```
mv hadoop-2.6.0 hadoop
```

- Создаем группу ***hadoop***:

```
sudo addgroup hadoop
```

- Добавляем пользователя ***hduser*** в группу ***hadoop***:

```
sudo usermod -a -G hadoop hduser
```

- Открываем файл ***/etc/sudoerc*** и добавляем туда нашего пользователя:

```
sudo nano hadoop /etc/sudoers
```

```
# User privilege specification
root    ALL=(ALL:ALL) ALL
hduser    ALL=(ALL:ALL) ALL
```

- Делаем нашего пользователя владельцем директории ***hadoop*** и всего ее содержимого:

```
sudo chown -R hduser:hadoop hadoop
```

- Добавляем в файл ***.bashrc*** переменные окружения hadoop:

```
nano .bashrc
```

```
export HADOOP_HOME=/home/hduser/hadoop
export HADOOP_CONF_DIR=/home/hduser/hadoop/etc/hadoop
export HADOOP_MAPRED_HOME=/home/hduser/hadoop
export HADOOP_COMMON_HOME=/home/hduser/hadoop
export HADOOP_HDFS_HOME=/home/hduser/hadoop
export YARN_HOME=/home/hduser/hadoop
export PATH=$PATH:/home/hduser/hadoop/bin
export PATH=$PATH:/home/hduser/hadoop/sbin
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_PREFIX/lib"
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
```

- Переподключаемся, чтобы подгрузились переменные окружения:

```
exit
ssh hduser@192.168.0.7
```

- Добавляем конфигурацию в ***core-site.xml***:

```
nano hadoop/etc/hadoop/core-site.xml
```

```
<configuration>
    <property>
        <name>fs.defaultFS</name>
        <value>hdfs://localhost:9000</value>
    </property>
    <property>
        <name>hadoop.proxyuser.hduser.hosts</name>
        <value>*</value>
    </property>
    <property>
        <name>hadoop.proxyuser.hduser.groups</name>
        <value>*</value>
    </property>
</configuration>
```

- Добавляем конфигурацию в ***hdfs-site.xml***:

```
nano hadoop/etc/hadoop/hdfs-site.xml
```

```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
</configuration>
```

- Форматируем файловую систему hadoop:

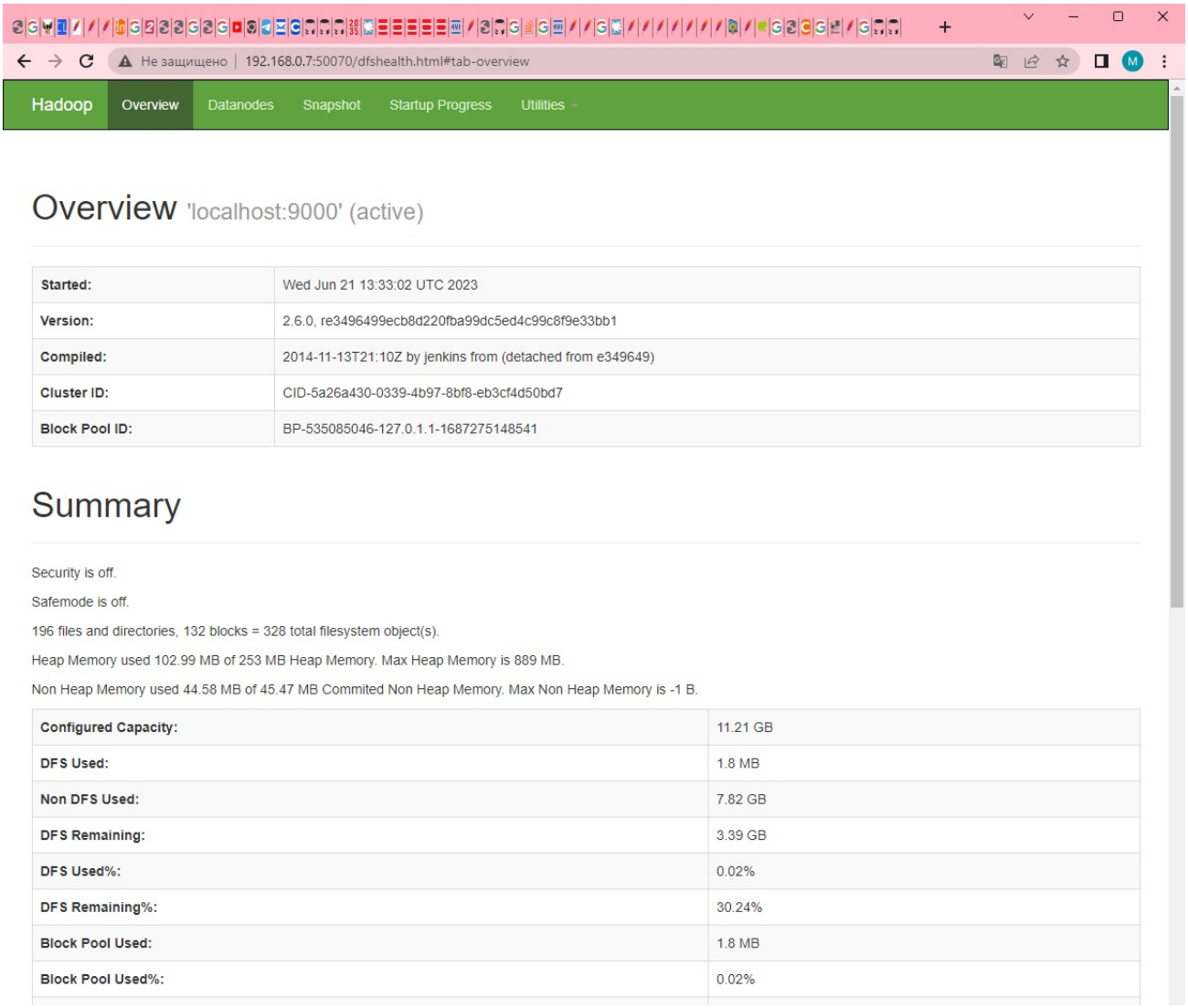
```
hdfs namenode -format
```

- Запускаем ***NameNode daemon*** и ***DataNode daemon***:

```
start-dfs.sh
```

- Проверяем доступность ***NameNode*** в веб-интерфейсе: http://192.168.0.7:50070/

***50070*** - порт по умолчанию для ***NameNode***




- Создаем директории в ***hdfs***, необходимые для выполнения заданий ***MapReduce***:

```
hdfs dfs -mkdir /user
hdfs dfs -mkdir /user/hduser
```

- Добавляем конфигурацию в ***mapred-site.xml***:

```
nano hadoop/etc/hadoop/mapred-site.xml
```

```
<configuration>
    <property>
        <name>mapreduce.framework.name</name>
        <value>yarn</value>
    </property>
</configuration>
```

- Добавляем конфигурацию в ***yarn-site.xml***:

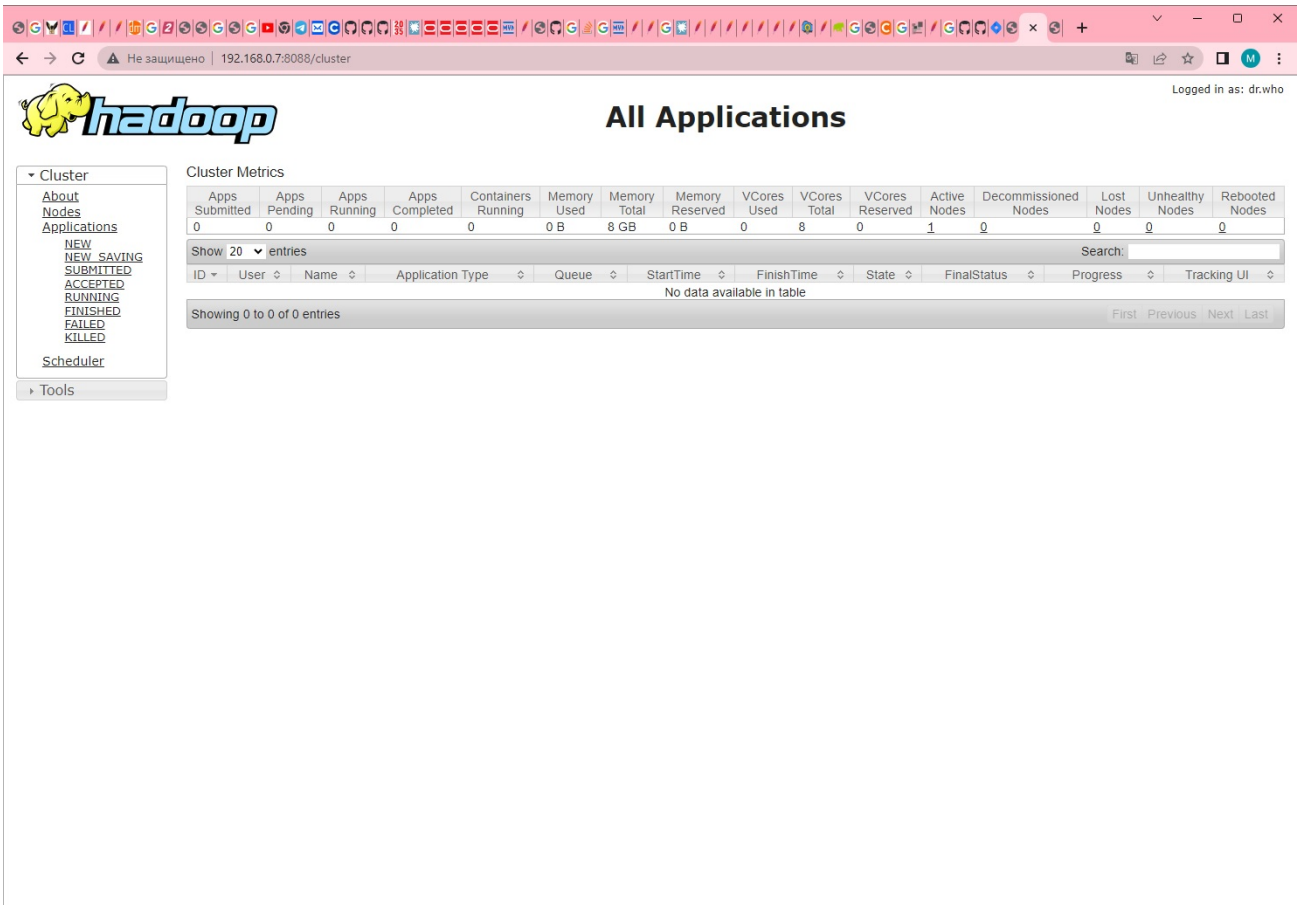
```
nano hadoop/etc/hadoop/yarn-site.xml
```

```
<configuration>
    <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
    </property>
</configuration>
```

- Запускаем ***ResourceManager daemon*** и ***NodeManager daemon***:

```
start-yarn.sh
```

- Проверяем доступность ***ResourceManager*** в веб-интерфейсе:
  http://192.168.0.7:8088/
  ***8088*** - порт по умолчанию для ***ResourceManager***




# Настройка Oozie-5.2.1

- Скачиваем архив ***Oozie-5.2.1*** в домашнюю директорию пользователя:

```
wget https://archive.apache.org/dist/oozie/5.2.1/oozie-5.2.1.tar.gz
```

- Разархивируем скачанный файл:

```
tar -xzvf oozie-5.2.1.tar.gz
```

- Проверяем версию ***hadoop*** в ***pom.xml***:

```
nano oozie-5.2.1/pom.xml
```

- Собираем ***Oozie*** с помощью скрипта ***mkdistro.sh***
  При сборке обнаружилось, что не удается собрать зависимости
  org.apache.hive.hcatalog:hive-hcatalog-server-extensions:jar:1.2.2 ->
  org.apache.hive.hcatalog:hive-hcatalog-core:jar:1.2.2 ->
  org.apache.hive:hive-cli:jar:1.2.2 ->
  org.apache.hive:hive-service:jar:1.2.2 ->
  org.apache.hive:hive-exec:jar:1.2.2 ->
  org.apache.calcite:calcite-core:jar:1.2.0-incubating ->
  org.pentaho:pentaho-aggdesigner-algorithm:jar:5.1.5-jhyde,
  т.к. страница сайта http://conjars.org/repo, где хранятся репозитории с зависимостями, недоступна более.
  Поэтому для успешной сборки используем флаг ***--fail-never***:

```
bin/mkdistro.sh -DskipTests --fail-never
```

```
[INFO] Apache Oozie Main ................................ SUCCESS [  6.765 s]
[INFO] Apache Oozie Fluent Job .......................... SUCCESS [  0.092 s]
[INFO] Apache Oozie Fluent Job API ...................... SUCCESS [01:56 min]
[INFO] Apache Oozie Client .............................. SUCCESS [ 17.893 s]
[INFO] Apache Oozie Share Lib Oozie ..................... SUCCESS [  2.775 s]
[INFO] Apache Oozie Share Lib HCatalog .................. FAILURE [02:15 min]
[INFO] Apache Oozie Share Lib Distcp .................... SUCCESS [  1.170 s]
[INFO] Apache Oozie Core ................................ SUCCESS [ 21.294 s]
[INFO] Apache Oozie Share Lib Streaming ................. SUCCESS [  1.934 s]
[INFO] Apache Oozie Share Lib Pig ....................... SUCCESS [  5.196 s]
[INFO] Apache Oozie Share Lib Git ....................... SUCCESS [  2.411 s]
[INFO] Apache Oozie Share Lib Hive ...................... SUCCESS [  3.535 s]
[INFO] Apache Oozie Share Lib Hive 2 .................... SUCCESS [  3.390 s]
[INFO] Apache Oozie Share Lib Sqoop ..................... SUCCESS [  1.960 s]
[INFO] Apache Oozie Examples ............................ SUCCESS [  3.378 s]
[INFO] Apache Oozie Share Lib Spark ..................... SUCCESS [  6.868 s]
[INFO] Apache Oozie Share Lib ........................... SUCCESS [ 16.039 s]
[INFO] Apache Oozie Docs ................................ SUCCESS [  0.867 s]
[INFO] Apache Oozie WebApp .............................. SUCCESS [  9.902 s]
[INFO] Apache Oozie Tools ............................... SUCCESS [  2.152 s]
[INFO] Apache Oozie MiniOozie ........................... SUCCESS [  1.544 s]
[INFO] Apache Oozie Fluent Job Client ................... SUCCESS [  0.955 s]
[INFO] Apache Oozie Server .............................. SUCCESS [  5.580 s]
[INFO] Apache Oozie Distro .............................. SUCCESS [ 33.023 s]
[INFO] Apache Oozie ZooKeeper Security Tests ............ SUCCESS [  4.083 s]
[INFO] ------------------------------------------------------------------------
[INFO] BUILD FAILURE
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  06:50 min
[INFO] Finished at: 2023-06-22T09:30:52Z
[INFO] ------------------------------------------------------------------------
```

- Создаем директорию ***libext*** по указанному пути и переходим в нее:

```
mkdir /home/hduser/oozie-5.2.1/distro/target/oozie-5.2.1-distro/oozie-5.2.1/libext
cd libext
```

- Скачиваем архив, необходимый для запуска ***Oozie Web Console***:

```
wget http://archive.cloudera.com/gplextras/misc/ext-2.2.zip
```

- Добавляем в файл ***.bashrc*** переменную ***OOZIE_HOME***:

```
nano /home/hduser/.bashrc
```

```
export HADOOP_HOME=/home/hduser/hadoop
export HADOOP_CONF_DIR=/home/hduser/hadoop/etc/hadoop
export HADOOP_MAPRED_HOME=/home/hduser/hadoop
export HADOOP_COMMON_HOME=/home/hduser/hadoop
export HADOOP_HDFS_HOME=/home/hduser/hadoop
export YARN_HOME=/home/hduser/hadoop
export PATH=$PATH:/home/hduser/hadoop/bin
export PATH=$PATH:/home/hduser/hadoop/sbin
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_PREFIX/lib"
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export OOZIE_HOME=/home/hduser/oozie-5.2.1/distro/target/oozie-5.2.1-distro/oozie-5.2.1
```

- Переподключаемся, чтобы подгрузились переменные окружения:

```
exit
ssh hduser@192.168.0.7
```

- Копируем библиотеки ***hadoop*** в директорию ***libext***:

```
cp $HADOOP_HOME/share/hadoop/common/*.jar $OOZIE_HOME/libext
cp $HADOOP_HOME/share/hadoop/common/lib/*.jar $OOZIE_HOME/libext
cp $HADOOP_HOME/share/hadoop/hdfs/*.jar $OOZIE_HOME/libext
cp $HADOOP_HOME/share/hadoop/hdfs/lib/*.jar $OOZIE_HOME/libext
cp $HADOOP_HOME/share/hadoop/mapreduce/*.jar $OOZIE_HOME/libext
cp $HADOOP_HOME/share/hadoop/mapreduce/lib/*.jar $OOZIE_HOME/libext
cp $HADOOP_HOME/share/hadoop/yarn/*.jar $OOZIE_HOME/libext
cp $HADOOP_HOME/share/hadoop/yarn/lib/*.jar $OOZIE_HOME/libext
```

- Добавляем конфигурацию в ***oozie-site.xml***:

```
nano $OOZIE_HOME/conf/oozie-site.xml
```

```
<configuration>
        <property>
                <name>oozie.service.HadoopAccessorService.hadoop.configurations</name>
                <value>*=/home/hduser/hadoop/etc/hadoop</value>
        </property>
        <property>
                <name>oozie.service.ProxyUserService.proxyuser.hduser.hosts</name>
                <value>*</value>
        </property>
        <property>
                <name>oozie.service.ProxyUserService.proxyuser.hduser.groups</name>
                <value>*</value>
        </property>
</configuration>
```

- Устанавливаем ***unzip***, необходимый для запуска скрипта ***oozie-setup.sh***:

```
sudo apt install unzip
```

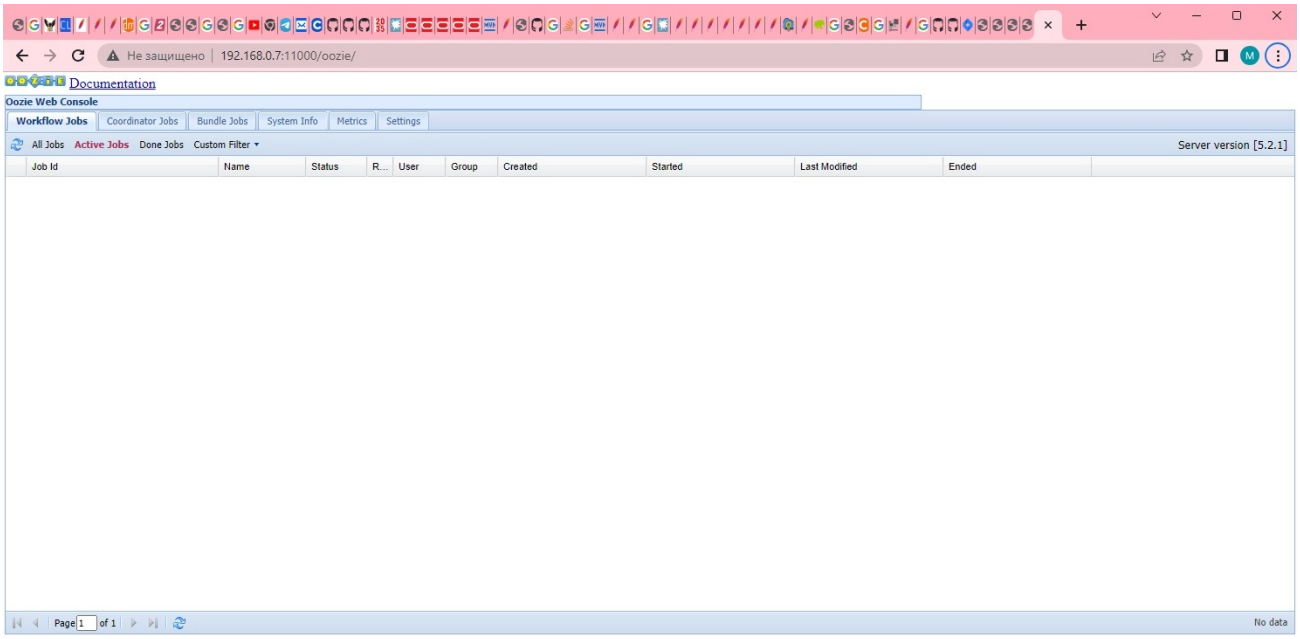
- Переходим в ***$OOZIE_HOME*** и запускаем команду для настройки ***Oozie***:

```
cd $OOZIE_HOME
bin/oozie-setup.sh sharelib create -fs /user/hduser/share/lib
```

- Запускаем ***Oozie***:

```
bin/oozied.sh start
```

- Проверяем доступность ***Oozie*** в веб-интерфейсе:
  http://192.168.0.7:11000/oozie
  ***11000*** - порт по умолчанию для ***Oozie***:




- Также можно проверить статус работы ***Oozie***:

```
./bin/oozie admin -oozie http://localhost:11000/oozie -status
```

```
System mode: NORMAL
```

# Запуск action в Oozie

- В директории ***$OOZIE_HOME*** распаковываем архив ***oozie-examples.tar.gz***:

```
tar -zxf oozie-examples.tar.gz
```

- Копируем директорию с разархивированным содержимым в ***hdfs***:

```
hadoop fs -put examples examples
```

- Выбираем пример для запуска, в нашем случае ***shell***, переходим в эту директорию и редактируем файл ***job.properties***, т.к. там указан неверный порт (***9000*** - порт по умолчанию):

```
nano $OOZIE_HOME/examples/apps/shell/job.properties
```

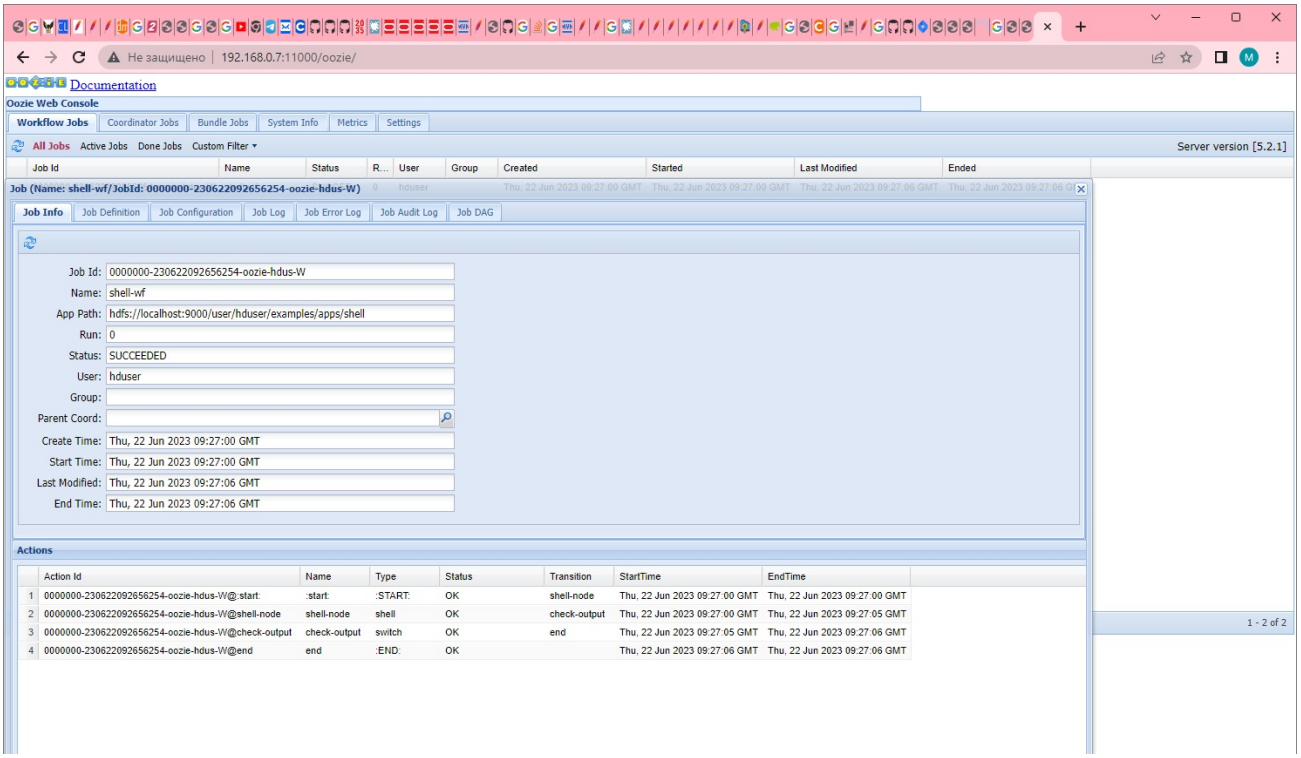
```
nameNode=hdfs://localhost:9000
resourceManager=localhost:8032
queueName=default
examplesRoot=examples

oozie.wf.application.path=${nameNode}/user/${user.name}/${examplesRoot}/apps/shell
```

- Запускаем ***Shell Action***:

```
bin/oozie job -oozie http://localhost:11000/oozie -config
$OOZIE_HOME/examples/apps/shell/job.properties -run
```

- Проверяем выполнение нашего ***Shell Action*** в веб-интерфейсе http://192.168.0.7:11000/
  ***11000*** - порт по умолчанию для ***Oozie***:




- Также можно проверить в консоле с помощью команды:

```
bin/oozie job -oozie http://localhost:11000/oozie -info 0000000-
230622092656254-oozie-hdus-W
```

```
Job ID : 0000000-230622092656254-oozie-hdus-W
------------------------------------------------------------------------------------------------------------------------------------
Workflow Name : shell-wf
App Path      : hdfs://localhost:9000/user/hduser/examples/apps/shell
Status        : SUCCEEDED
Run           : 0
User          : hduser
Group         : -
Created       : 2023-06-22 09:27 GMT
Started       : 2023-06-22 09:27 GMT
Last Modified : 2023-06-22 09:27 GMT
Ended         : 2023-06-22 09:27 GMT
CoordAction ID: -

Actions
------------------------------------------------------------------------------------------------------------------------------------
ID                                                                            Status    Ext ID                 Ext Status Err Code
------------------------------------------------------------------------------------------------------------------------------------
0000000-230622092656254-oozie-hdus-W@:start:                                  OK        -                      OK         -
------------------------------------------------------------------------------------------------------------------------------------
0000000-230622092656254-oozie-hdus-W@shell-node                               OK        application_1687424915715_0001SUCCEEDED  -
------------------------------------------------------------------------------------------------------------------------------------
0000000-230622092656254-oozie-hdus-W@check-output                             OK        -                      end        -
------------------------------------------------------------------------------------------------------------------------------------
0000000-230622092656254-oozie-hdus-W@end                                      OK        -                      OK         -
------------------------------------------------------------------------------------------------------------------------------------
```