|  |  |
| --- | --- |
|  | **Работа со Spark на Hadoop-кластере BigData Team** |

# 

# **Доступы к сервисам Spark**

1. Заходим на сервер:

ssh <USER>@virtual-client.bigdatateam.ru -L 8088:virtual-master:8088

1. Открываем в браузере: <http://localhost:8088/>

# **Запуск Jupyter + Spark в распределенном режиме**

Для запуска нужно ввести следующую команду (т.к. команда будет работать пока работает Jupyter-ноутбук, её лучше запустить в фоновом режиме или tmux).

|  |
| --- |
| PYSPARK\_DRIVER\_PYTHON=jupyter PYSPARK\_PYTHON=/usr/bin/python3 PYSPARK\_DRIVER\_PYTHON\_OPTS='notebook --ip="\*" --port=<PORT> --NotebookApp.token="TOKEN" --no-browser' pyspark2 --master=yarn --executor-memory=3GB |

- PORT - “100” + 2 последние цифры вашего логина из таблицы <https://docs.google.com/spreadsheets/d/1kJVmDUUZppQtQHP5hDeP0jLP_VmukkYHbwhrKMMo8-o/edit?usp=sharing>

- TOKEN - токен, который нужно будет ввести для входа в Jupyter (любая строка),

После этого заходите в браузере на [**http://virtual-client.bigdatateam.ru**](http://virtual-client.bigdatateam.ru)**:<PORT>** и вводите TOKEN.

# **3. Работа с GraphFrames**

Чтоб подключить Graphframes, нужно сделать 2 вещи:

1. В параметры pyspark2 при запуске ноутбука добавить:

--jars /opt/graphframes/graphframes-assembly-0.7.0-spark2.3-SNAPSHOT.jar

Т.е. получим такую команду:

|  |
| --- |
| PYSPARK\_DRIVER\_PYTHON=jupyter PYSPARK\_PYTHON=/usr/bin/python3 PYSPARK\_DRIVER\_PYTHON\_OPTS='notebook --ip="\*" --port=<PORT> --NotebookApp.token="TOKEN" --no-browser' pyspark2 --jars /opt/graphframes/graphframes-assembly-0.7.0-spark2.3-SNAPSHOT.jar --master=yarn --executor-memory=3GB |

Пример:

|  |
| --- |
| PYSPARK\_DRIVER\_PYTHON=jupyter PYSPARK\_PYTHON=/usr/bin/python3 PYSPARK\_DRIVER\_PYTHON\_OPTS='notebook --ip="\*" --port=10005 --NotebookApp.token="TOKEN" --no-browser' pyspark2 --jars /opt/graphframes/graphframes-assembly-0.7.0-spark2.3-SNAPSHOT.jar --master=yarn --executor-memory=3GB |

1. В самой верхней строке ноутбука прописать:

|  |
| --- |
| sc.addPyFile('/opt/graphframes/graphframes-assembly-0.7.0-spark2.3-SNAPSHOT.jar') |

При этом SparkContext подключать не надо, переменная sc существует в ноутбуке по умолчанию. В этом легко убедиться, просто выполнив в ноутбуке: `sc`.