Инфраструктура и доступы

Вы получаете доступ к инфраструктуре проекта через Telegram-бот: https://t.me/YanPracticumBot

▼ Из чего состоит инфраструктура для проекта

1. Виртуальная машина, на ней запущен Docker (как в спринте «Актуализация модели данных»). Внутри:
    a. Postgres
    b. Airflow
    c. Metabase
    d. VSC
2. Кластер Dataproc из 3 машин — мастера и двух воркеров:
    a. HDFS, Hadoop, Spark
    b. Вам необходимо будет запустить на мастере Jupyter Notebook

▼ Где взять данные

https://data.ijklmn.xyz/events/ — скачать на виртуальную машину с помощью wget/curl.

▼ Чтобы запустить инфраструктуру,

нажмите в боте:

Запустить инфраструктуру > **Модуль 8: Хакатон (проект)**

Достаточно сделать это одному человеку в команде.

▼ Что выдаёт бот в ответе

1. IP адрес виртуальной машины с Docker:
   - для подключения можно использовать `ssh -i <provided_key> yc-user@<IP>`
2. IP адрес мастер-ноды Dataproc-кластера:
   - для подключения смотрите инструкцию ниже
3. Intranet URL мастер-ноды Dataproc-кластера:
   - через неё можно смотреть UI сервисов после открытия тоннеля: YARN History UI, Spark UI
4. Приватная часть SSH-ключа для подключения к обеим компонентам.

▼ Подключение к виртуальной машине и Docker:

▼ Использование терминала

1. Сохраните полученный от бота файл, например, по пути `~/secret_key`
2. Назначьте ему правильные права: `chmod 0400 ~/secret_key`
3. Используйте ключ для подключения к ВМ: `ssh -i ~/secret_key yc-user@<IP>`
   ВЫ ПОДКЛЮЧЕНЫ К ВИРТУАЛЬНОЙ МАШИНЕ, дальше вы можете подключиться к терминалу в Docker, если есть такая необходимость.
4. Подключение к Docker:
   a. Найдите ID или имя (name) контейнера для подключения: `docker ps`
   b. Подключитесь в интерактивный терминал: `docker exec -it <cont_id> bash`

▼ Использование интерфейсов сервисов

1. Airflow: <IP>:3000/airflow/
   логин: AirflowAdmin
   пароль: airflow_pass

2. VSCode <IP>:3000/vsc/, но есть риск, что он будет работать плохо. **РЕКОМЕНДУЕТСЯ** писать код локально и загружать на машину, используя SCP или репозиторий Git (git clone, git push/pull)
3. Metabase: <IP>:3000/metabase/
4. Postgres: <IP>:5432
   логин: jovyan
   пароль: jovyan

▼ Подключение к кластеру DataProc и запуск Jupyter Notebook

▼ Установка соединения с интранетом:

Открыть тоннель (прокси) до мастер-ноды, возьмите SSH-ключ, полученный через Telegram-бот, и IP адрес мастер-ноды (пользователь Ubuntu): `ssh -i ~/secret_key -ND 8157 ubuntu@<IP>`

- **ДЛЯ WINDOWS:** посмотрите инструкцию во введению к спринту «Организация DataLake» по туннелированию через PuTTy:

Настройка Proxy

Вам понадобится сделать два шага.

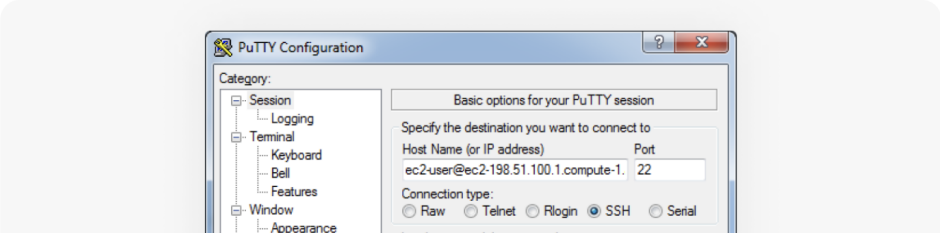
1. Открыть SSH-тоннель. Таким образом, вы получите доступ к узлам в интранете.

   Для Linux или Mac OC

   - Введите в терминал команду `ssh -i <private_key.file> -ND 8157 <username>@<IP>`, заменив текст в угловых скобках на данные, которые вам выдаст Telegram-бот. Эта команда откроет тоннель на локальном порту 8157 до узла <IP>.
   - Не останавливайте команду и не закрывайте окно терминала: оно поддерживает тоннель в рабочем состоянии.

   Для Windows

   - Скачайте и запустите утилиту PuTTY.



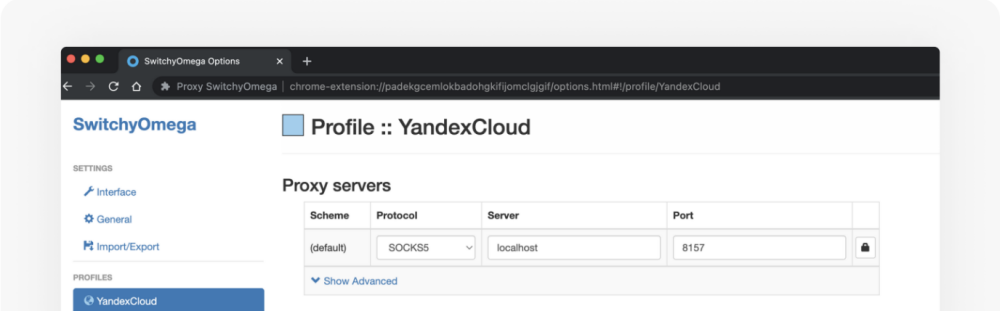
1. Настроить расширение в браузере для использования только что запущенной Proxy:

2. Настроить расширение в браузере для использования тоннеля (proxy). Для Яндекс Браузера и Google Chrome можно использовать Proxy SwitchyOmega, для Firefox — FoxyProxy.

   Инструкция для Google Chrome:

   - Установите Proxy SwitchyOmega.
   - В расширении:
     - Откройте настройки, кликнув на значок расширения в правом верхнем углу и выбрав “Options”.
     - Создайте новый профиль. Назовите его, например, YandexCloud.
     - В списке серверов выберите:

       Protocol — SOCKS5, Server — localhost, Port - 8157.



2. Проверить доступность, открыв интранет-ссылку на мастер-ноду в браузере, например: <MASTER_NODE_URL>:8088

▼ Подключение по SSH и запуск Jupyter Notebook

1. Открытие SSH-подключения: `ssh -i ~/secret_key ubuntu@<MASTER_NODE_IP>`
   - **ДЛЯ WINDOWS**: создайте SSH-подключение, используя PuTTY (придётся сконвертировать ключ в puttygen для использования с PuTTY).
2. Запуск Jupyter Notebook:
   a. В терминале запустите программу `screen` (виртуальный терминал)
   b. «В скрине» запустите Jupyter Notebook командой:
      `jupyter notebook --ip 0.0.0.0 --no-browser --port=8889`
   c. После запуска в консоли отобразится путь, включающий в себя токен. Перейдите по этому адресу в любом удобном для вас

браузере, не забудьте в пути изменить IP (0.0.0.0) на IP-адрес мастер-ноды.

1. Отключитесь от сессии в `screen` сочетанием клавиш: `CTRL+A+D`