ETL: автоматизация подготовки данных

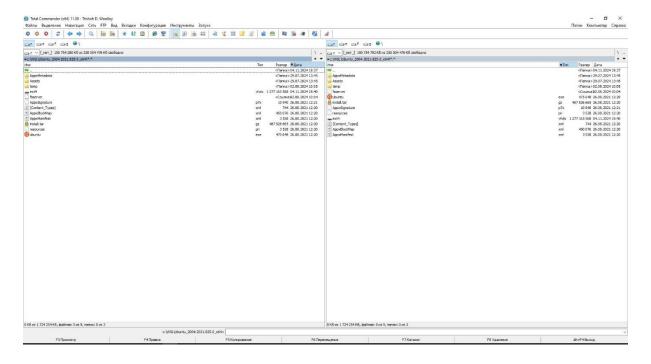
Урок 5. Обзор возможностей Airflow, установка и настройка

Задание

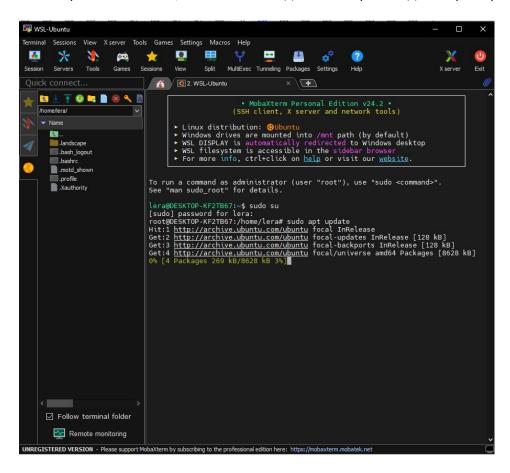
Установите AirFlow как показано на семинаре, соберите список команд установки, скриншоты GUI и работы AirFlow из мобы. Соберите все данные в один pdf файл..

Установка AirFlow на платформу Windows 10

- 1. Для начала установки AirFlow на платформу Windows 10 нужно включить встроенную виртуальную машину Linux.
- 1.1. В системе Windows 10 зайдем в "Панель управления" -> "Программы и компоненты" -> "Включение и отключение компонентов "Windows".
- нужно включить либо удостовериться что включены следующие компоненты (проставить галочки):
 - Платформа для виртуальной машины;
 - Подсистема Windows для Linux;
- перезагрузить компьютер.
- 1.2. В PowerShell введем команду:
 - wsl --set-default-version 2 (запустим вторую версию wsl)
- 1.3. Установим и запустим Ubuntu (из Ubuntu_2004.2021.825.0_x64.appx):

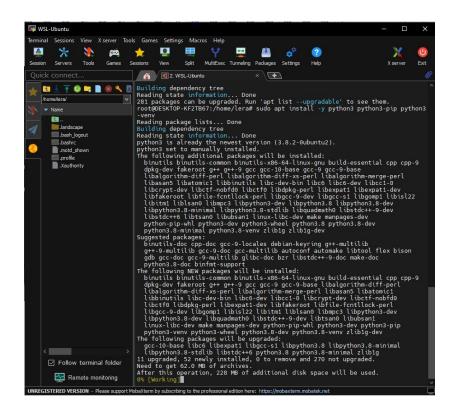


1.4. Закрываем Ubuntu, все остальные действия производим через программу MobaXterm:

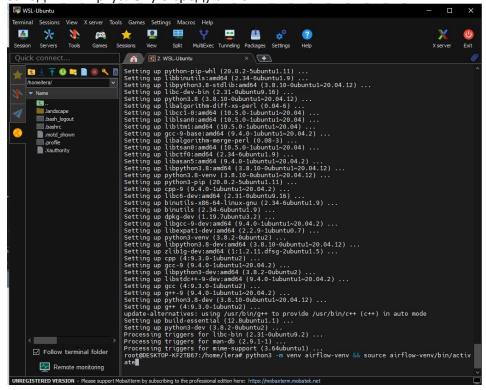


- 2.2 sudo su получаем права доступа
- 2.3 sudo apt update получаем пакеты обновлений для Linux
- 2.4 sudo add-apt-repository ppa:deadsnakes/ppa подключаем нужный нам репозиторий
- 2.5 sudo apt install -y python3 python3-pip python3-venv команда устанавливает Python, pip и venv последней версии доступной в поключённом репозитории. python3: Это интерпретатор языка программирования Python версии 3.х. Он позволяет запускать и выполнять Python-скрипты и приложения. python3-pip: Это инструмент управления пакетами для Python. Он используется для установки, обновления и управления сторонними библиотеками и модулями Python. python3-venv:

Этот пакет предоставляет виртуальное окружение Python, которое позволяет изолировать проекты Python друг от друга. Виртуальные окружения полезны для управления зависимостями проекта и предотвращения конфликтов между версиями пакетов. - sudo apt install -y python3.8 python3.8-pip python3.8-venv пример команды если нам нужна конкретная версия Python, в данном случае Python 3.8



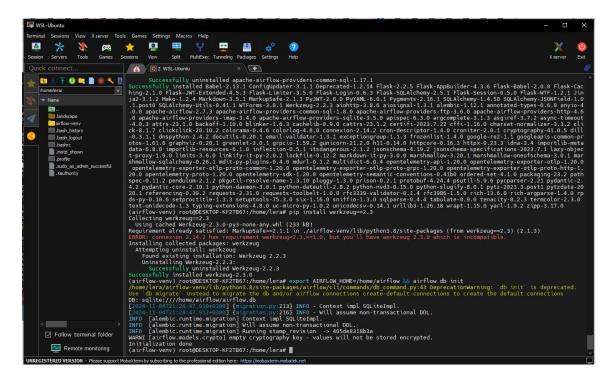
- 3. Установка программы AirFlow
- 3.1 python3 -m venv airflow-venv && source airflow-venv/bin/activate данной командой мы заходим в виртуальную среду airflow



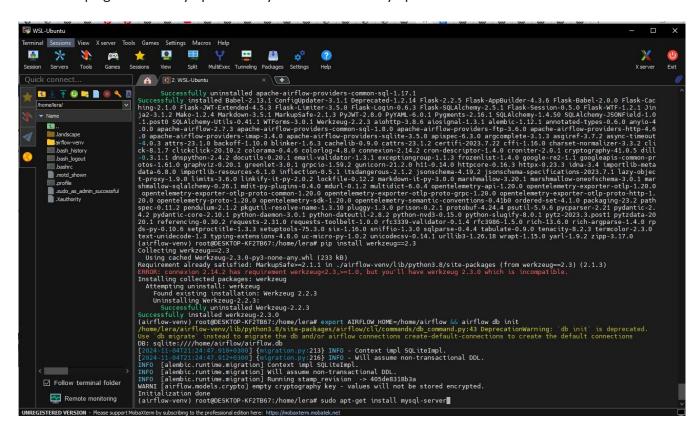
3.2 pip install apache-airflow==2.7.3 --constraint

"<a href="https://raw.githubusercontent.com/apache/airflow/constraints-2.7.3/constraints-3.8.txt" --force-reinstall --upgrade данной командой мы устанавливаем airflow

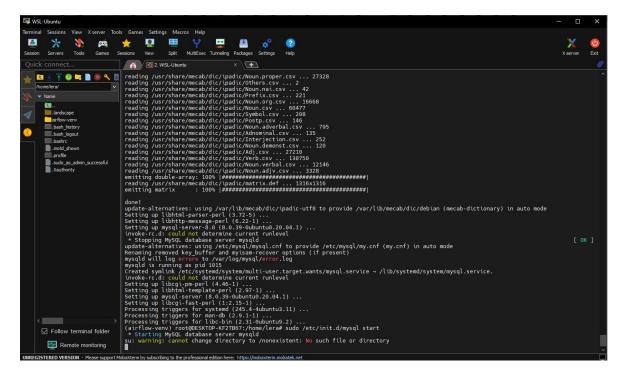
3.3 export AIRFLOW_HOME=/home/airflow && airflow db init - делаем пробный запуск airflow



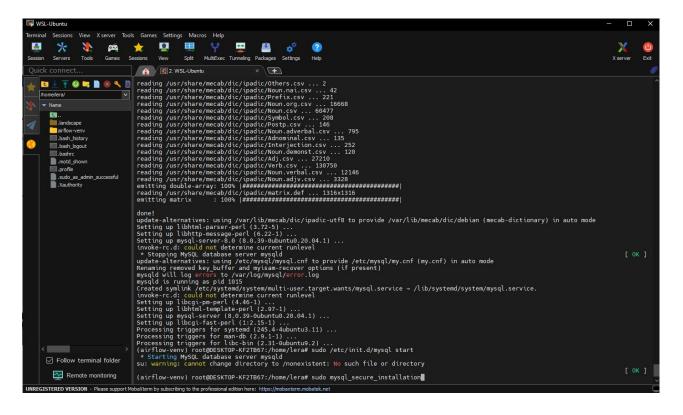
- 4. Далее мы устанавливаем Mysql:
- 4.1 sudo apt-get install mysql-server устанавливаем Mysql



4.2 sudo /etc/init.d/mysql start - запускаем Mysql

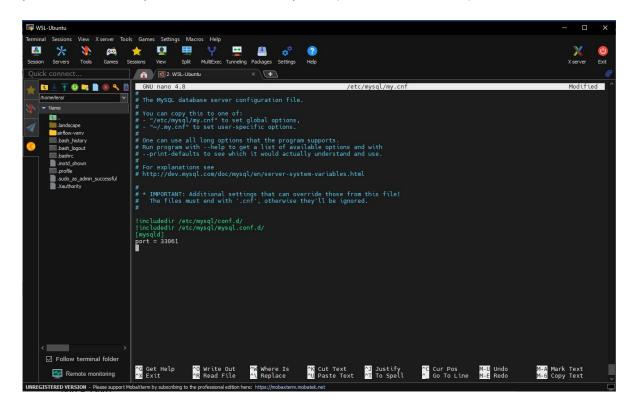


- 4.3 sudo mysql_secure_installation запускаем настройку сервиса везде ставим нет кроме последнего пункта:
- VALIDATE PASSWORD COMPONENT... ставим n
- By default, a MySQL installation has an anonymous user... ставим n (no)
- Normally, root should only be allowed to connect from 'localhost'... ставим n (no)
- By default, MySQL comes with a database named 'test' that... ставим n (no)
- Reloading the privilege tables will ensure that all changes made so far will take effect immediately anyone can access... ставим у (yes).

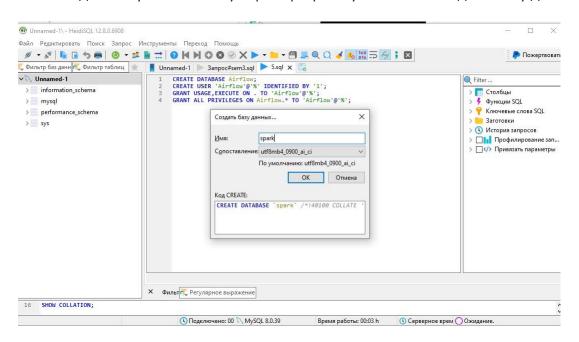


- 4.4. sudo mysql -u root -p заходим в программу Mysql
- 4.5. ALTER USER 'root'@'localhost' IDENTIFIED WITH mysql_native_password BY '1'; строка выше установит пароль "1", затем выходим командой exit
- 4.6.sudo nano /etc/mysql/my.cnf данный файл надо подредактировать добавив строчки: [mysqld]

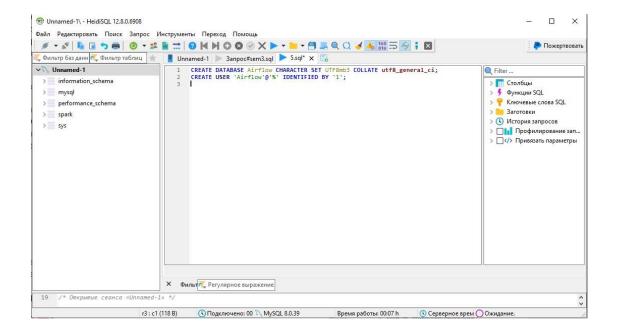
port = 33061 и сохраняем изменения в файле (Clrl+X -> Y -> Enter)

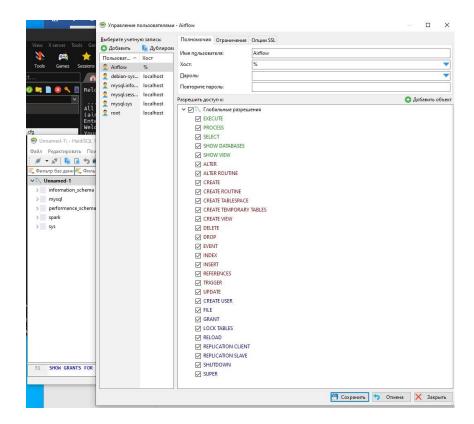


- 4.7. sudo service mysql restart перезапускаем Mysql
- 4.8. заходим в MySQL оболочку через программу HeidiSQL и создаем базу данных "spark":

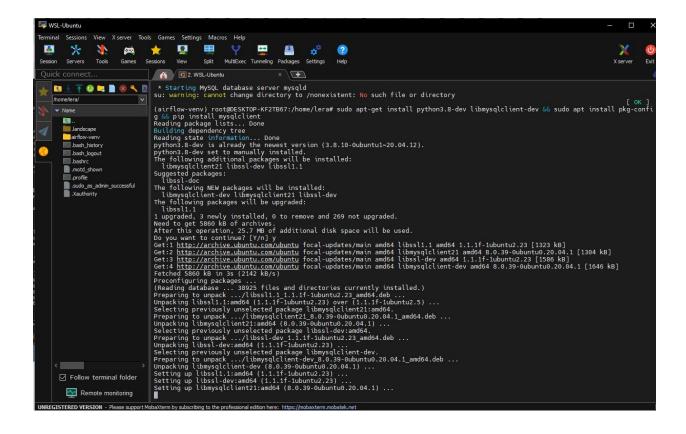


4.9 Сделаем SQL-запрос: CREATE DATABASE Airflow CHARACTER SET UTF8mb3 COLLATE utf8_general_ci; CREATE USER 'Airflow'@'%' IDENTIFIED BY '1'; - в результате появится база данных Airflow, затем заходим в пользователей и выдаем пользователю Airflow все права доступа.

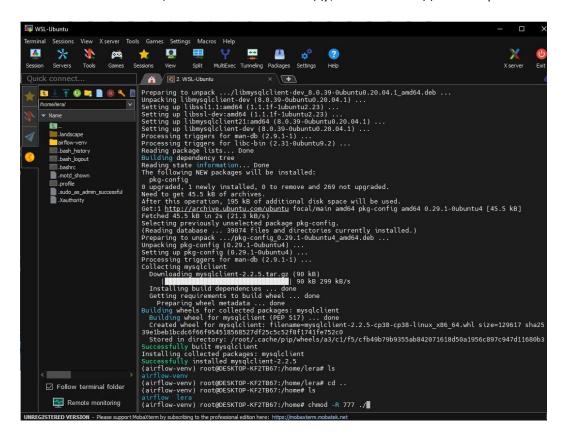




- 5 Продолжаем работать с установкой программы AirFlow
- 5.1 apt-get install python3.8-dev libmysqlclient-dev && sudo apt install pkg-config && pip install mysqlclient устанавливаем библиотеки:
 - pip install python3.8 установит дополнительно питон нужной версии
 - pip install pkg-config установит дополнительную библиотеку
 - pip install mysqlclient установит нужный mysqlclient

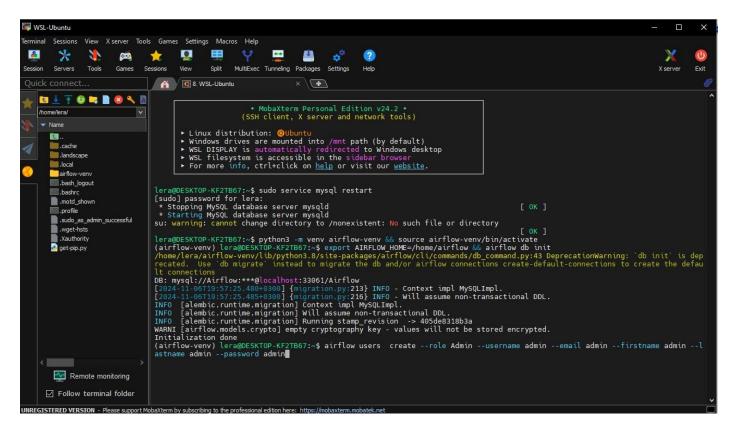


5.2. chmod -R 777 ./ - выполняем команду, затем снова делаем пробный запуск airflow

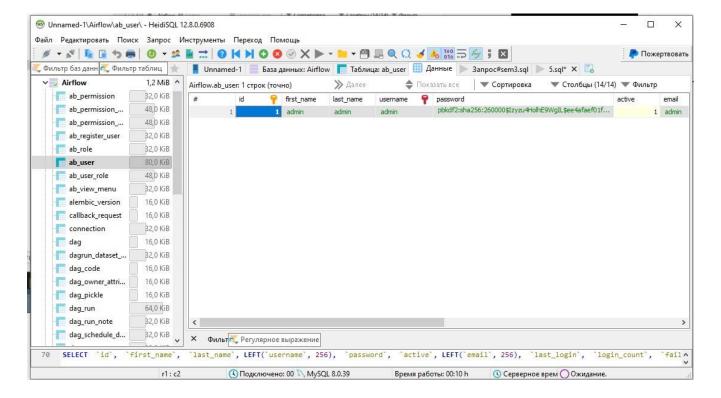


- 5.2.1. Появилась папка "airflow" там есть файл airflow.cfg который, мы должны заменить нашим файлом airflow.cfg
- 5.2.2. В каталоге "/home/airflow/" создаем папку "dags"

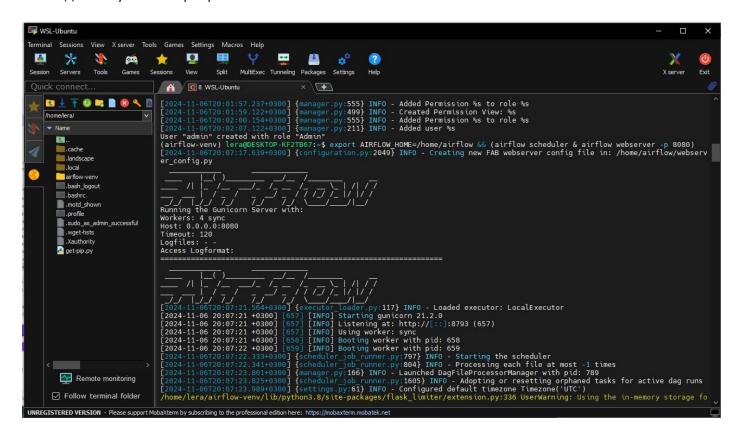
- 5.2.3. export AIRFLOW_HOME=/home/airflow && airflow db init делаем повторную иницализацию. Выдает ошибку.
- 5.3 Выключаем wsl (wsl shutdown.bat) и затем запускаем заново.
- 5.4 Снова инициализируем виртуальную среду и инициализируем airflow



5.5. airflow users create --role Admin --username admin --email admin --firstname admin --lastname admin --password admin - данной командой создаём пользователя для airflow



5.6 export AIRFLOW_HOME=/home/airflow && (airflow scheduler & airflow webserver -p 8080) - данной командой запускаем сервер airflow



Для подключения к серверу airflow в адресной строке браузера вводим localhost:8080 в результате чего появляется админ сайт airflow - логин/пароль: admin

Мы установили и запустили AirFlow в Windows 10 с помощью WSL

