ETL: автоматизация подготовки данных

Урок 6. Обзор возможностей Airflow, установка и настройка

Задание

Установить спарк как показано на семинаре:

- Для этого переместите папку spark в home.

- Дайте права командой chmod -R 777 ./

- nano ~/.bashrc

- export SPARK\_HOME=/home/spark && export

PATH=$PATH:$SPARK\_HOME/bin:$SPARK\_HOME/sbin

- source ~/.bashrc

- sudo apt-get install openjdk-8-jdkpip

- Указанные библиотеки нужно также установить и в виртуальную среду: python3-m venv airflow venv && source airflow venv /bin/activate

- pip install pyspark==3.2.4 && pip install pandas==1.5.3 && pip install SQLAlchemy==1.4.46

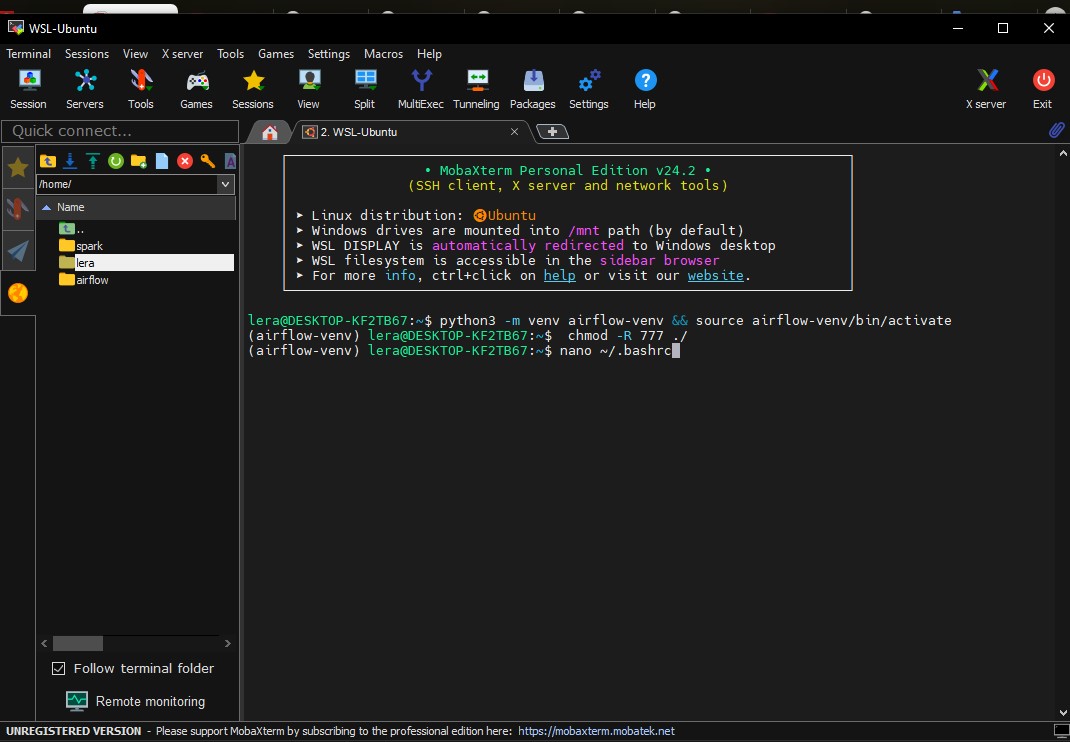
Используйте ДЗ которые вы мне высылали для 3-4 семинара. Запустите данные

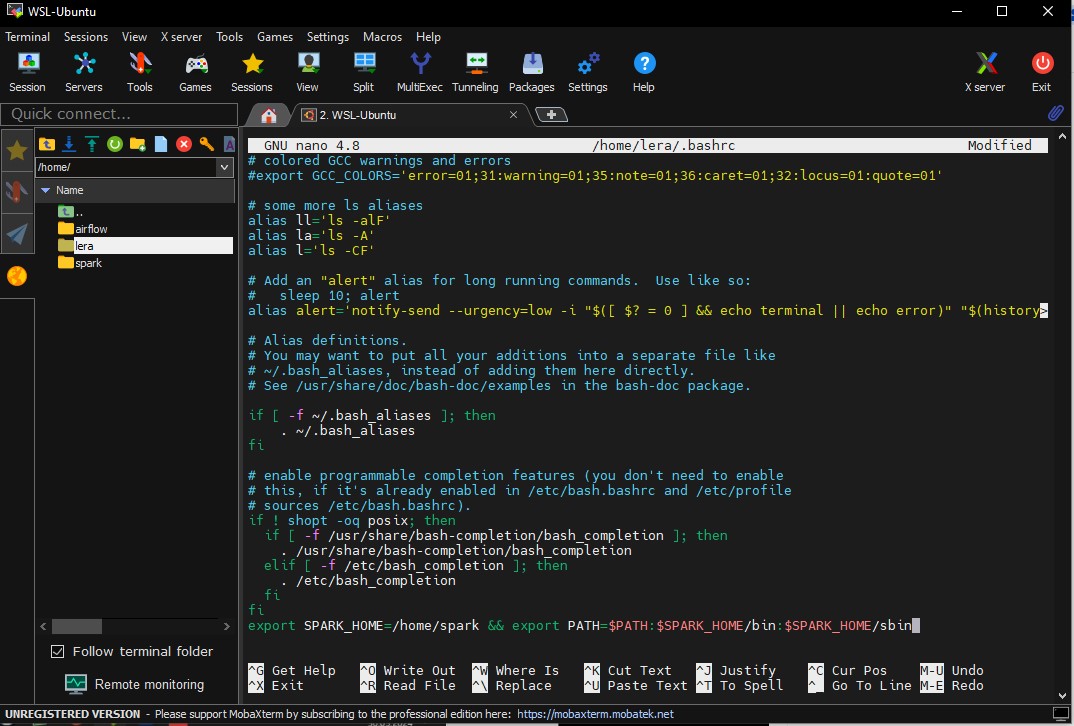
задачи ПОСЛЕДОВАТЕЛЬНО, одну за другой в аирфлоу. Пришлите мне скриншоты

выполненных задач в аирфлоу, логов аирфлоу, скриншоты что у вас записались

таблицы в БД mysql на WSL. По возможности доработайте код чтобы изображение с

линии платежей генерировалось в указанную директорию. Скриншоты соберите в pdf.





sudo apt-get install openjdk-8-jdkpip (не сработало, чтобы запустилось пришлось делать так:

sudo add-apt-repository ppa:openjdk-r/ppa

sudo apt-get update

sudo apt-get install openjdk-8-jdk)

