Курсовой проект для студентов учебного центра Neoflex по направлению «Data-Engineering (BDS)»

Предисловие

Данная курсовая работа состоит из нескольких задач. Первая треть посвящена работе с реляционными БД (Oracle или PostgreSQL), остальные относятся к стэку BigData, а именно к технологии Spark. Но при этом у всех заданий есть общий принцип – работа с данными, в большинстве с помощью SQL. И важно заметить, что хорошо уметь делать выборки и трансформации данных – это лишь полдела, так же нужно уметь выстраивать полноценный автономный ETL-процесс. То есть ещё важно развивать навык архитектурного / системного мышления. Это очень ценный навык в продуктовой и корпоративной разработке.

Легенда

Некий банк выполнил миграцию с одной БД на другую. В ходе миграции были частично сломаны некоторые ETL-процессы и частично утеряны некоторые данные. Вам предстоит это исправить.

Цели курсовой работы:

1. Развить и продемонстрировать навык самостоятельного проектирования и разработки ETL-систем;
2. Научиться устанавливать и настраивать реляционную СУБД (Oracle или PostgreSQL);
3. Изучить основные принципы создания и работы с процедурами в Oracle или PostgreSQL;
4. Научиться устанавливать и настраивать фрэймворк Spark внутри ОС «Ubuntu»;
5. Потренероваться в написании SQL скриптов и полноценных процессах с помощью одного из 3-х языков (Python, Scala, Java);
6. Научиться кратко и содержательно презентовать разработанные решения.

Перечень технологий:

* Реляционная СУБД: Oracle 12+ или PostgreSQL 13+;
* Среда SQL разработки (одна из): SQL Develop; pgAdmin; DBeaver; HeidiSQL; FrogSQL;
* Python 3.7+, Java 8 и, по желанию, Scala;
* Talend 7.x (по желанию);
* Virtual Box;
* Ubuntu 19+;
* Spark / PySpark;
* Jupyter notebook (опционально).

Примечания:

* Для комфортного выполнения зданий вам понадобиться компьютер со следующими минимальными требованиями: win 7 / win 10 / Linux; 4 ГБ ОЗУ; 4-х ядерный процессор.  
  Если у вы не обладаете устройством с подобными характеристиками – обратитесь к кураторам или организаторам УЦ. Вероятно, вам выделят компьютер в офисе (если вы в филиальном городе);
* В первую очередь вам понадобиться реляционная СУБД. Это может быть Oracle или PostgreSQL. В идеале установить Oracle 21XE (локально), но можно и версии пониже. Вам нужно зарегистрироваться на сайте oracle.com (лучше под VPN), указав выдуманное название компании и «дальнюю» страну (это формальность) и почту (лучше gmail.com);  
  <https://www.oracle.com/cis/database/technologies/xe-downloads.html>  
  Видео инструкция по установке ораклового сервера:  
  <https://www.youtube.com/watch?v=-BJgLNzSssg&list=LL&index=10>  
  Текстовая статья по установке:  
  <https://ithelpspb.blogspot.com/2020/12/oracle-database-19c-windows-server-2019.html>  
    
  Если возникнут трудности с Oracle, то можно установить PostgreSQL 13 и pgAdmin;  
  <https://www.enterprisedb.com/downloads/postgres-postgresql-downloads>
* Python лучше установить не ниже версии 3.7.  
  Java лучше установить не ниже 8-й версии, с более высокими версиями бывают проблемы при интерграции с различными инструментами, но это лишь предположение;
* Не забудьте после установки Java / Python настроить системные переменные;
* (\*) Если вы имеете опыт работы с Linux-системами (например Ubuntu) и сможете установить туда Oracle или PostgreSQL и работать с БД – это будет плюсом для вас.