Лекция 6. Практическое задание

Курс: разработчик ХД

Лекция 6: Практическое задание (1 из 3)

1. Реализуйте SQL-запросы наполнения пользовательских **витрин** данных. Требования:
   - Наполнение витрины маржинальности выполняется на одну отчетную **дату** (один день).
   - Наполнение витрины матрицы продаж выполняется **целиком**.
   - ETL-процесс должен поддерживать **многократный** запуск с сохранением корректного результата на приемнике (без задубливания данных и т.п.).

- **Подсказка**: реализовать наполнение **матрицы** продаж можно двумя основными способами:
  - через множество SQL-операторов **CASE** по отчетной дате в одном SELECT запросе и последующей агрегации колонок месячных продаж;
  - через множество SELECT запросов с фильтрами по отчетной дате, результаты которых объединяются конструкцией **UNION ALL** и последующей агрегацией колонок месячных продаж.
- Оператор CASE более требователен к CPU-ресурсам, UNION ALL потребует многократного чтения таблиц DDS. Реализуйте загрузку любым подходящим для вас способом.
- **Обратите внимание**: наполнение витрины маржинальности на всю глубину истории можно реализовать разово одним запросом, но запрос будет работать несколько часов и требовать огромного количества ресурсов! Лучше витрину наполнять последовательно, порциями за 1 месяц!

Лекция 6: Практическое задание (2 из 3)

Для решения второй части практического задания вам потребуется:

- **редактор** кода (например VSCode, PyCharm, etc.);
- установленный **интерпретатор** Python 3.10 и выше;
- клиент для работы с **ssh**, например, WinSCP и PuTTY (можно работать из консоли).

Настройте новое подключение к серверу практических заданий Airflow.

- Подключение выполнять с включенным VPN!
- **Адрес** сервера: http://10.10.144.45:8080/
- **Логин**: studentXX (номер, предоставленный куратором).
- **Пароль**: studentXX (номер, предоставленный куратором).

Обратите внимание: чтобы ваш DAG появился в web UI Airflow необходимо подключиться по ssh к серверу 10.10.144.45 и **скопировать** файл в каталог /opt/airflow/dags.

Подсказка: при написании кода желательно придерживаться стандарта pep8.

Лекция 6: Практическое задание (3 из 3)

2. Реализуйте **полную цепочку** ETL-процесса по наполнению и **обновлению** данных **витрин** в Apache Airflow: Источники -> Staging -> (ODS, если есть) -> DDS -> витрины.
   - Вызывайте созданные ранее пользовательские **функции** Greenplum в Apache Airflow.
   - Выстройте вызов функций в правильной **последовательности**. Загрузка последующего слоя не должна начинаться до окончания загрузки всех таблиц предыдущего слоя!
   - При запуске расчета должна быть возможность указать **дату**, за которую производится расчет.
   - Для простоты наполнение всех слоев и расчет витрин можно реализовать в **одном** DAG файле.

Выполните ETL-процесс в Airflow и убедитесь в появлении новых/обновленных данных в DDS и витринах (из DBeaver).

- Результаты практического Задания 1 лекции 6 принимаются в файлах форматов .sql или .txt.
- Результаты практического Задания 2 лекции 6 принимаются в виде ссылок на ваш DAG в Airflow.

Если вы не знаете языка Python и не в состоянии выполнить Задание 2, реализуйте ETL-цепочку в виде одной функции Greenplum, в ней пропишите последовательность вызова всех остальных функций.
За такую реализацию **-3 балла** от максимально возможной оценки!