Итоговая презентация

по курсу «Инженер данных»

Выполнил: Ляш Олег Иванович

Название и общее описание проекта.

- Проект №5
- На основе данных о поездах такси в г. Нью-Йорк необходимо подготовить таблицу **parquet**. В этой таблице необходимо показать процентное соотношение количества пассажиров на каждый день. Также необходимо добавить максимальную и минимальную стоимости поездок.

Цели проекта с описанием бизнес-задачи и требованиями

- «Необходимо, используя таблицу поездок для каждого дня рассчитать процент поездок по количеству человек в машине (без пассажиров, 1, 2,3,4 и более пассажиров). По итогу должна получиться таблица (parquet) с колонками date, percentage_zero, percentage_1p, percentage_2p, percentage_3p, percentage_4p_plus. Технологический стек – sql,scala (что-то одно)».
- Также добавить столбцы к предыдущим результатам с самой дорогой и самой дешевой поездкой для каждой группы.

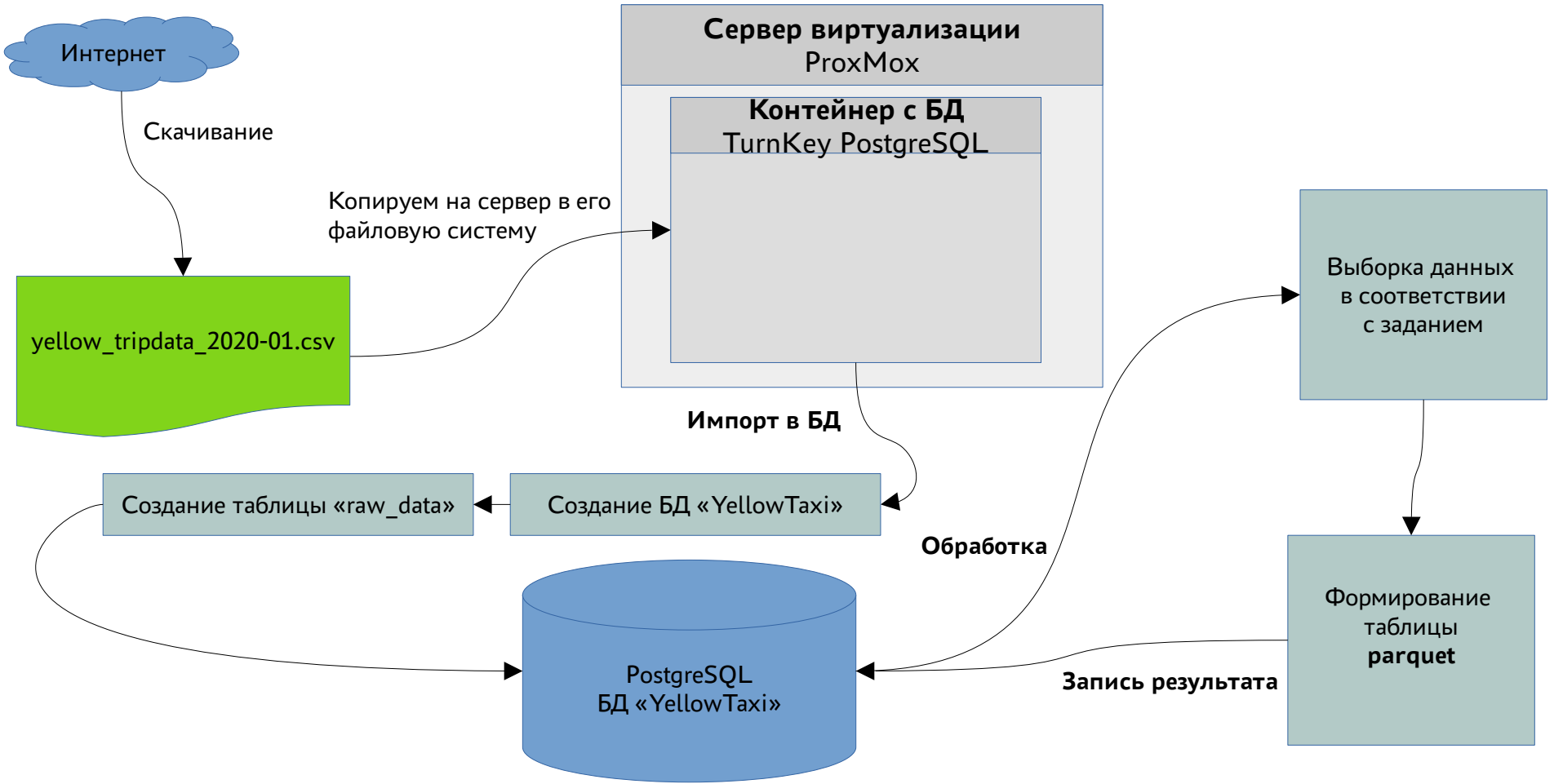
План реализации

- Анализ исходных данных
- Создание БД и таблицы для размещения не обработанных данных
- Импорт исходных данны в БД
- Реализация запроса для построения требуемой таблицы parquet
- Оформление отчёта

Используемые технологии с обоснованием

- Сервер виртуализации: Proxmox Virtual Environment (PVE) - хорошо зарекомендовавшее себя решение для виртуализации компонентов инфраструктуры предприятия. Достоинства: opensource и бесплатно.
- дистрибутив TurnKey-PostgreSQL — готовый контейнер с установленной СУБД. Достоинства: opensource, бесплатно, готово к работе в PVE.
- Сервер баз данных: PostgreSQL — производительная СУБД. Достоинства: opensource, бесплатно, большое русскоязычное сообщество.
- Инструмент для работы с сервером баз данных: DataGrip — привычный инструмент для работы с БД. Достоинства: бесплатно в учебных целях.
- Оболочка для подключения к серверу: git-bash — привычная среда для работы с консолью в ОС Windows. Достоинства: бесплатно.

Схемы/архитектуры с обоснованием



Результаты разработки

- Создана БД и таблица для не обработанных данных
- Импортированы полученные из сети данные
- Создан SQL запрос для выборки требуемых данных
- Сформирована таблица parquet с результатами выборки

Выводы

- В целом поставленная задача выполнена
- Возможности выбранных компонентов достаточны для решения поставленной задачи
- В будущем целесообразно рассмотреть решение задачи с использованием возможностей python и модулей: pandas, numpy и т.п.