Описание задания: В качестве исходных данных можно использовать любой публичный датасет с разными объектами, либо же сгенерировать свой: например, с помощью модуля Faker. Для вдохновения можно использовать одну из придуманных нами тем:

- Агентство путешествий во времени (сущности — машины времени, бронирование, предлагаемые временные периоды, клиенты...)
- Менеджмент космических экспедиций (сущности — корабли, экипаж, миссии, грузы, вооружение...)
- Виртуальные питомцы (сущности — питомцы, игрушки, активности, пользователи, игровые события...)
- Коллекционные карточные игры (сущности — карты, аукцион, колоды, фракции...)
- Коллективный блог о видеоиграх (сущности — статьи, отзывы, рейтинги по разным параметрам...)

Постановка задачи содержит следующие опорные пункты:

1. описание бизнес-процессов, происходящих в выбранном домене, вариант структуры исходных данных. Постановка задач глазами бизнеса — какую область они хотят исследовать (можно краткие описания желаемых витрин и отчетов по ним);
2. указание ожидаемых видов и структур данных на входе, а также их объем и период сбора (пакетный/потоковый);
3. преобразования для оценки и повышения Data Quality — возможно использование специализированного инструмента, а также подходов, о которых говорилось на лекциях (приведение null, типов);
4. архитектура слоев хранения данных — сырое хранилище (если нужно), структурированное (если нужно), витрины и т.п.;
5. инструмент(ы) для передачи данных между сервисами (ETL), примеры преобразований над данными — агрегирование, PIVOT, фильтрацию и т.д.;
6. описание подходов по автоматизации процессов, либо части из них — скрипты cron/Apache Airflow/Apache NiFi/etc.;
7. предоставление данных заказчику — витрины, дашборды, BI-инструментарий;

Замечание. Данные можно брать рабочие (сделав mock на критичных для отображения вещах) или из открытых источников.

Ответом на задание является: презентация по пунктам с архитектурной схемой, обоснованием выбора конкретных технических решений, структур и форматов хранения данных на разных слоях и этапах, возможно, с примерами кода (Python/SQL/и т.п.). В идеале — примерный план по требуемым серверам для расчета бюджета. У вас будет возможность в течение 10- 15 минут провести техническую презентацию своего проекта по этим слайдам и получить фидбэк

Инструменты, которые могут пригодиться для выполнения: *Python, Airflow, NiFi, Kafka, dbt, Clickhouse, Postgres, MongoDB, Metabase.* Также можно использовать инструменты, не включенные в программу, например, Superset для визуализации, RabbitMQ для передачи данных.

Критерии, по которым будет оцениваться задание: быстродействие и надежность систем, соответствие выбранных технических решений заявленным задачам. Оценивается обоснованное использование инструментов.