**Дипломная работа по профессии Инженер данных**

**Цель:** составить документацию процессов ETL на основе предложенного датасета

**Этапы выполнения дипломной работы**

1. Обработайте и проанализируйте данные
2. Сформируйте нормализованную схему данных (NDS)
3. Сформируйте состав таблиц фактов и измерений (DDS)
4. Сформируйте ETL-процессы: для заливки данных в NDS и для создания витрин
5. Сформируйте набор метрик и дашбордов на их основе
6. Оформите результаты, сформулируйте выводы

**Задание:**

[Данные для выполнения дипломного задания](https://www.kaggle.com/aungpyaeap/supermarket-sales?select=supermarket_sales+-+Sheet1.csv)

1, Вам необходимо разработать и задокументировать ETL-процессы заливки данных в хранилище, состоящее из слоёв:

* NDS - нормализованное хранилище и DDS - схема звезда;
* Data Quality - опционально, будет большим преимуществом в вашей работе

2. на основании DDS построить в Табло дашборды

***Рекомендации при выполнении работы:***

1, ETL процессы можно делать:

* с помощью Pentaho;
* с помощью Python (pandas) + SQL;

.  
2, датасет:

* предложен вам в CSV формате выше;
* сбор данных вы также можете сделать из сторонних API, это станет вашим преимуществом;

3, дополнительно вы можете сделать оркестровку с помощью Airflow;

4, опционально можно сделать отдельный слой метаданных в хранилище, а также дашборды на основании данных из этого слоя, где будет отображаться кол-во прогрузок и их статусы;

**Результат:**

* дашборды
* задокументированная схема хранилища данных
* документированная схема ETL-процессов

**Формат выполнения:** дипломная работа носит комплексный подход, поэтому рекомендуем подготовить к защите воркбуки Табло, ERR-диаграммы для схемы хранилища + ktr/kjb файлы с ETL-процессами или py-файлы с DAG Airflow