# **Финальный проект**

Проект нужно выполнить самостоятельно.

Он должен быть залит на Github (залить проект одним коммитом нельзя – будет проверяться история).

### **Цель проекта**

Применить полученные навыки и построить полноценную Data Platform с нуля.

### **Критерии оценки**

1. Техническая полнота проекта. Все компоненты Data Platform должны быть корректно реализованы.
2. Эффективность системы. Обработка данных не должна занимать большое количество времени.
3. Качество данных. Не должно быть ложных данных, дубликатов и пр.
4. Полноценность структуры Хранилища Данных. Все данные из источников должны присутствовать в Хранилище Данных и корректно организованы.
5. Качество кода.

# **Требования**

### **Источники**

* API
  + out of stock
* Данные из таблиц PostgreSQL
  + orders
  + products
  + departments
  + aisles
  + clients
  + stores
  + store\_types
  + location\_areas

### **Составные компоненты приложения**

* Источники данных:
  + PostgreSQL
  + API
* Data Lake:
  + Apache Hadoop HDFS
* Хранилище данных:
  + Greenplum
* ETL:
  + Apache Spark
* Scheduler & ETL Management:
  + Apache Airflow

### **ETL**

Extract:

* Для каждого источника должен быть создан отдельный DAG в Airflow
* Каждая таблица PostgreSQL должна обрабатываться в отдельном Airflow операторе внутри одного DAG
* Данные должны храниться в Bronze Layer Data Lake

Transform:

* Отдельный(е) Airflow DAG, которые должны провалидировать данные в Bronze. Провести проверку данных и перенести их в Silver
* Данные в Silver должны быть организованы

Load:

* Загрузить данные из Silver прямо в Хранилище Данных, без использования промежуточного хранилища

ETL нужно проектировать с учетом, что джобы будут выполняться ежедневно в заранее заданное время.

Все credentials и пр. информацию необходимо хранить в Airflow Connections.

Использование Airflow XCom допустимо только для мета информации.

### **DWH**

* Хранилище данных должно быть реализовано в виде схемы Звезды
* Должно быть, как минимум, две таблицы фактов