Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное
учреждение высшего образования
«Самарский национальный исследовательский университет имени академика
С.П. Королева»
Институт информатики, математики и электроники Факультет информатики
Кафедра технической кибернетики

Отчет по лабораторной работе №1
по дисциплине «Инженерия данных»

Тема: «Знакомство с основными инструментами построения пайплайнов: Apache Airflow и Apache Nifi»

Направление подготовки: 01.04.02 Прикладная математика и информатика:
Магистерская программа «Наука о данных»

Профиль «Системы искусственного интеллекта»

Выполнила Шаина М. М.,
студентка группы 6231 – 010402D

Самара 2023

В рамках данной лабораторной работы предлагается построить простейший пайплайн (рисунок 1), собирающий воедино данные из нескольких файлов, обрабатывающий их и сохраняющий результат в no-sql базу данных.

Цель лабораторной работы – знакомство с основными инструментами построения пайплайнов: Apache Airflow и Apache Nifi.

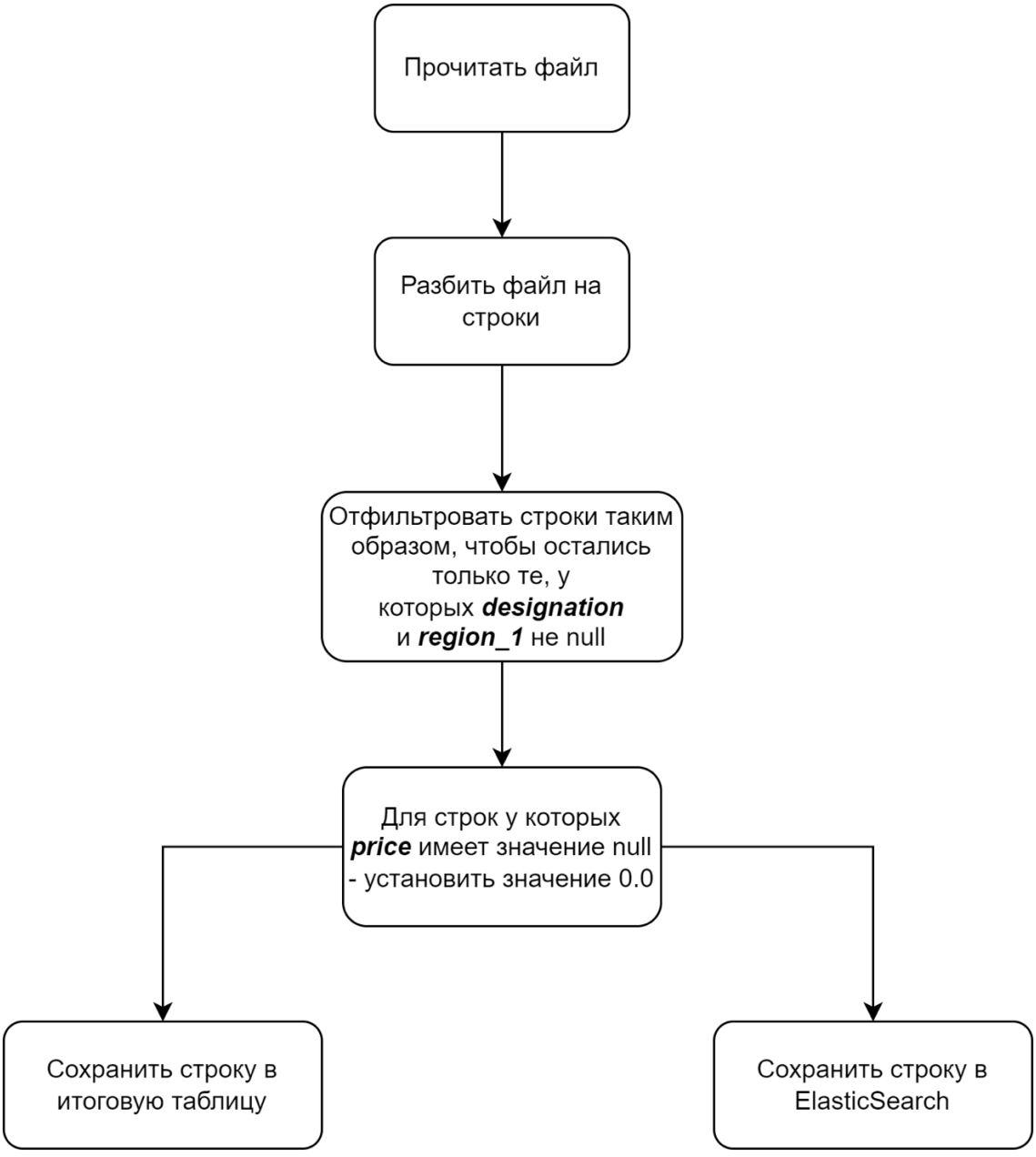

Рисунок 1 — Схема описывающая пайплайн, который необходимо построить в рамках лабораторной работы

Также необходимо построить гистограмму стоимости напитка к баллам средствами Kibana. Результат построения гистограммы представлен на рисунке 4.

Для построения пайплайна рассматриваются инструменты, такие как ElasticSearch, Kibana, Apache Airflow и Apache Nifi.

В качествен данных - будем использовать набор из нескольких CSV файлов, полученных из набора данных wine-review.

Описание основных шагов:

1. Настройка программ для выполнения лабораторной работы.
2. Выполнения пайплайна в Apache Nifi.
3. Выполнение пайплайна в Apache Airflow.

Работа выполнялась в Virtual Studio Code со следующими расширениями:

1. ms-python.python
2. ms-toolsai.jupyter
3. ms-vscode-remote.vscode-remote-extensionpack
4. ms-azuretools.vscode-docker

1 Выполнения пайплайна в Apache Nifi

Перед начало работы в Apache Nifi csv файлы необходимо перенести в папку, которой Nifi имеет доступ, например, в папку `nifi/data/lab_1/input` (рисунок 2, см. ниже).

Для построения пайплайна в Apache Nifi использовались следующие процессоры:

1. GetFile
2. SplitRecord
3. QueryRecord
4. UpdateRecord
5. MergeContent
6. PutFile

7. PutElasticsearcRecord

Более подробное описание процессоров приведено в таблице 1.

Таблица 1 — Описание процессоров Apache Nifi

Processor	**Name**	**Property**	**Value**
GetFile	GetFile	Input Directory	/opt/nifi/nifi-current/data/lab_1/input
		File Filter	[^\.].*
SplitRecord	SplitRecord	Record Reader	CSVReader
		Record Writer	CSVRecordSetWriter
		Record Per Split	100 000
QueryRecord	Drop designation&region_1 IsNull Filter	isnull	SELECT * FROM FLOWFILE WHERE designation IS NOT NULL AND region_1 IS NOT NULL
UpdateRecord	Update price FromNulltoZero	/price	${field.value:ifEmpty('0.0'):toNumber()}
MergeConten	MergeConten	Delimiter Strategy	Text
		Header	id,country,description,designation…
PutFile	PutFile	Output Directory	/opt/nifi/nifi-current/data/lab_1/output

Для объединения данных из разных ветвей в Funnel выбраны два процессора: PutFile и PutElasticsearcRecord

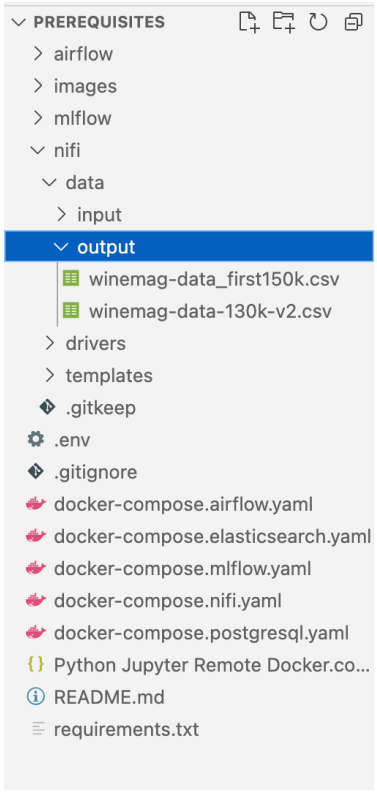

Рисунок 2 — Результат копирования csv файлов в VS Code

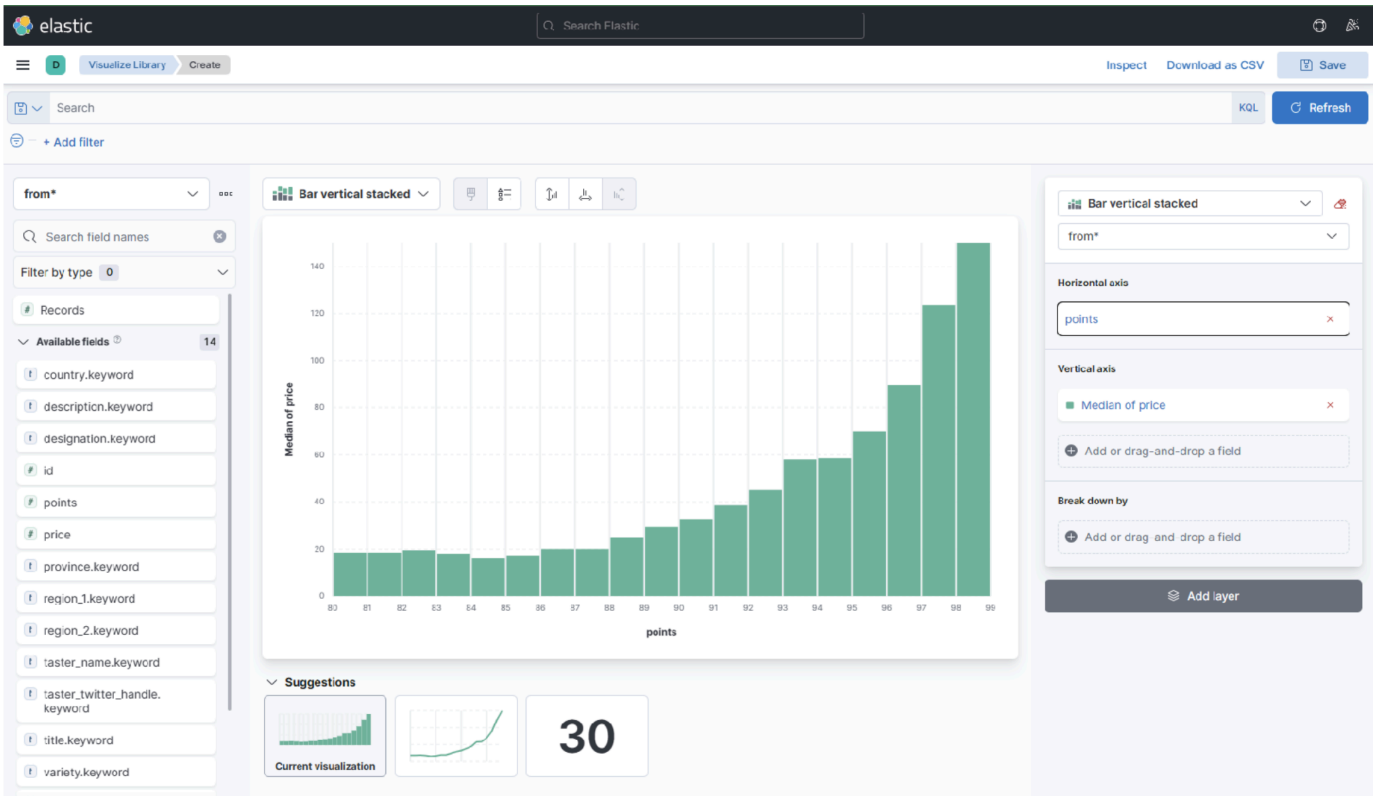

Рисунок 3 — Гистограмма зависимости цены от очков

2 Выполнения пайплайна в Apache Airflow

Перед началом работы необходимо авторизироваться, с предоставленным логином и паролем: http://localhost:18080/. После выполнения этого шага можно начать работу.

Используя Directed acyclic graph (DAG), реализуем пайплан для обработки и индексации CSV-файлов в Elasticsearch в Apache Airflow. Основная обработка в Python – preprocessing: удаление строк с пропущенными значениями в столбцах «designation» и «region_1», а также заполнение пропущенных значений в столбце «price» нулями. Код программы на языке Python представлен в файле airflow-lab1.py. Граф внутри Apache Airflow представлен на рисунке 4.

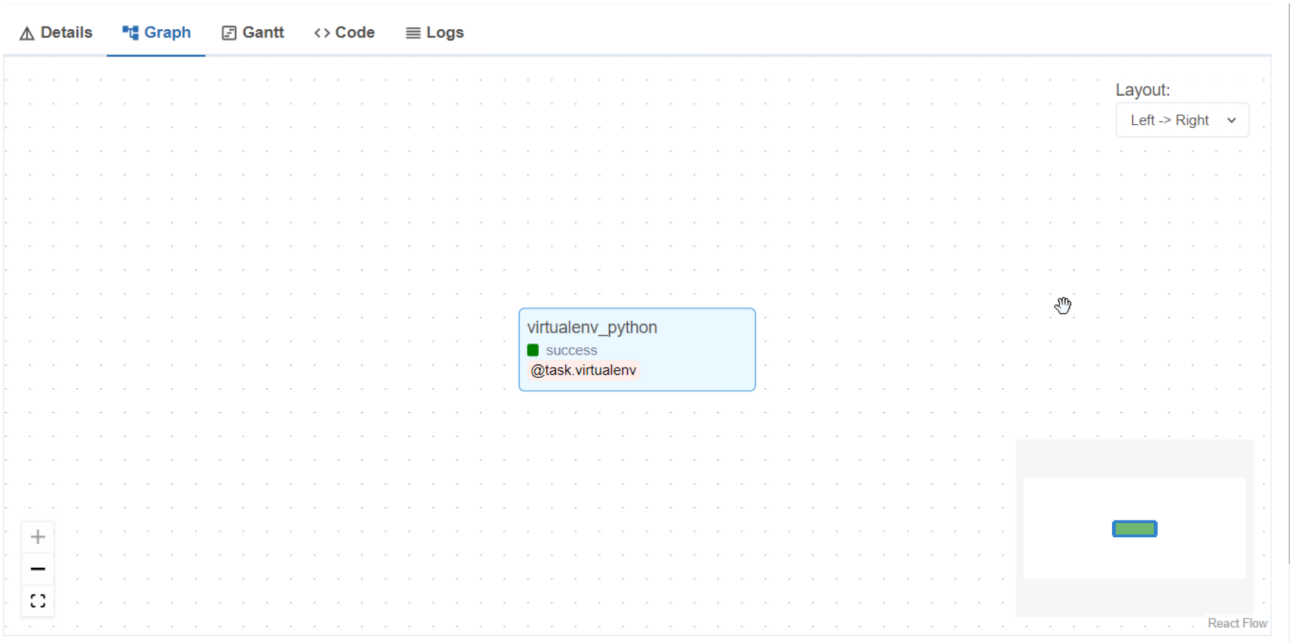

Рисунок 4 — DAG внутри Apache Airflow