

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное
образовательное учреждение высшего образования
«Самарский национальный исследовательский университет
имени академика С.П. Королева»
(Самарский университет)

Институт информатики и кибернетики
Кафедра технической кибернетики

Отчет по лабораторной работе №1

Дисциплина: «Инженерия данных»

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

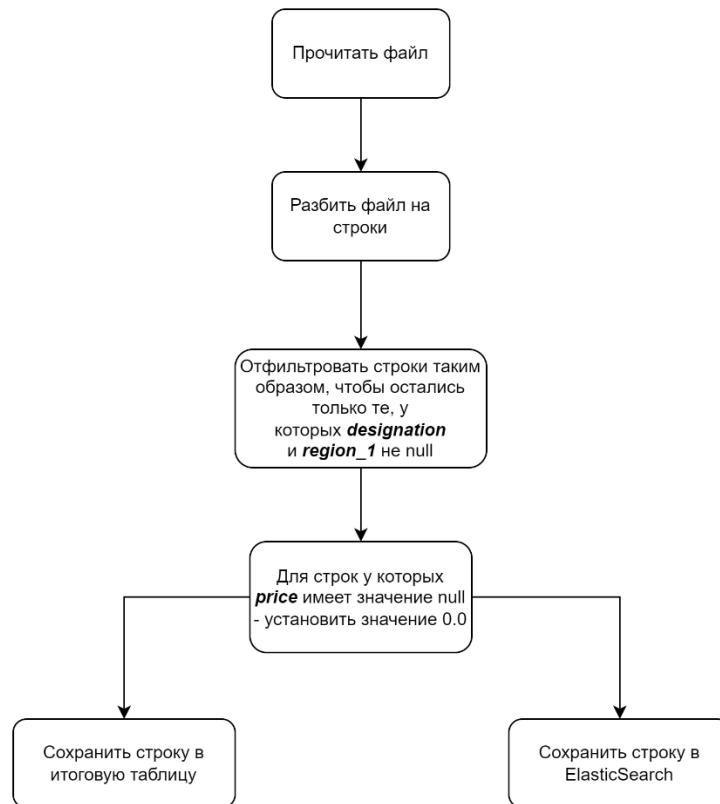
Выполнил: Каспаров И. А.

Группа: 6232-010402D

Самара 2024

Задание на лабораторную работу

Схема, описывающая пайплайн, который необходимо построить в рамках лабораторной работы:



Данный пайплайн должен быть построен *дважды*: один раз с использованием Apache Nifi и второй раз с использованием Apache Airflow.

Также средствами Kibana построить гистограмму зависимости стоимости напитка от баллов, поставленных дегустаторами.

Под сохранением в итоговую таблицу подразумевается объединение всех строк, прошедших через пайплайн, в единую таблицу формата .csv

1. Apache Airflow

Перед разворачиванием контейнеров поднимем соединение с Docker, при помощи команды:

\$ docker network create data-engineering-labs-network

Для работы контейнеров скачаем и установим для этого необходимые компоненты при помощи команды:

\$ docker compose -f docker-compose.airflow.yaml up airflow-init

```
C:\Users\ikasp\vscode\labs\prerequisites> docker network create data-engineering-labs-network
71c29960d2b57447273a18f21c932b08d8f4bc14174d0d0602188de2bb0da
C:\Users\ikasp\vscode\labs\prerequisites> docker compose -f docker-compose.airflow.yaml up airflow-init
time="2024-12-02T21:54:30+00:00" level=warning msg="C:\Users\ikasp\vscode\labs\prerequisites\docker-compose.airflow.yaml: the attribute 'version' is obsolete, it will be ignored, please remove it to avoid potential confusion"

redis Pulled
6f0d1bf7b091 Download complete
928f5d0d0807 Download complete
8f805c30417c Download complete
92efcc0809f Download complete
5e0ba07561c Download complete
74c736080471 Download complete
4f4ef700e5f4 Already exists
postgres Pulled
6f0d1bf7b091 Download complete
834023cac7f3 Download complete
933401227a69 Download complete
4ac3010d8cc7 Download complete
7853a8591a0b Download complete
2042909e73a0 Download complete
3ff1313a4ff Download complete
9e091ba0d5c2 Download complete
55e7ac6818e2 Download complete
47e09f64904c Download complete
f504c055a057 Download complete
c0b28346072 Download complete
f54e77052ec4 Download complete
c6a5d0807311 Download complete
[*] Building 5.4s (8/8) FINISHED

[+] Building 5.4s (8/8) FINISHED                                docker:desktop-linux
-> transferring dockerfile: 100B
-> [airflow-init internal] load build definition from Dockerfile
-> [airflow-init internal] load metadata for docker.io/apache/airflow:2.7.0
-> [airflow-init] authn: docker.io/pull token for registry-1.docker.io
-> [airflow-init internal] load .dockerignore
-> transferring context: 3B
-> [airflow-init 1/2] FROM docker.io/apache/airflow:2.7.0@sha256:7a1d5453d511e6d20093c34dbaa0417a3218a7162d01d4c51056d0d201d4e
-> resolve docker.io/apache/airflow:2.7.0@sha256:7a1d5453d511e6d20093c34dbaa0417a3218a7162d01d4c51056d0d201d4e
-> CACHED [airflow-init 2/2] RUN pip install --no-cache-dir airflow
-> [airflow-init] exporting to image
-> exporting layers
-> exporting manifest sha256:161a3c1297a9e1d06771a0d0c93107c180e072e180a9ffc8201400991770d3
-> exporting config sha256:4c7179567f89916d959d21091a0b15493d3d0e0d74302593594e415d080da
-> exporting attestation manifest sha256:187938a70e2777c121f936c94e912135a0a01577cc1cc44939dcf30ba3
-> exporting manifest list sha256:7993b0c1f1ef18400000a34487004f7240bc47784393d4d1009212b0eac1a
-> naming to docker.io/library/airflow-airflow-init:latest
-> unpacking to docker.io/library/airflow-airflow-init:latest
-> [airflow-init] resolving provenance for metadata file
Volume "airflow-postgres-db-volume" created
Container airflow-postgres-1 created
Container airflow-redis-1 created
Container airflow-airflow-init-1 created
attaching to airflow-init-1
airflow-init-1 | The container is run as root user. For security, consider using a regular user account.
airflow-init-1 |
```

Для разворачивания airflow используем команду:

\$ docker compose -f docker-compose.airflow.yaml up --build -d

```
C:\Users\ikasp\vscode\labs\prerequisites> docker compose -f docker-compose.airflow.yaml up airflow-init
time="2024-12-02T21:54:30+00:00" level=warning msg="C:\Users\ikasp\vscode\labs\prerequisites\docker-compose.airflow.yaml: the attribute 'version' is obsolete, it will be ignored, please remove it to avoid potential confusion"

redis Pulled
6f0d1bf7b091 Download complete
928f5d0d0807 Download complete
8f805c30417c Download complete
92efcc0809f Download complete
5e0ba07561c Download complete
74c736080471 Download complete
4f4ef700e5f4 Already exists
postgres Pulled
6f0d1bf7b091 Download complete
834023cac7f3 Download complete
933401227a69 Download complete
4ac3010d8cc7 Download complete
7853a8591a0b Download complete
2042909e73a0 Download complete
3ff1313a4ff Download complete
9e091ba0d5c2 Download complete
55e7ac6818e2 Download complete
47e09f64904c Download complete
f504c055a057 Download complete
c0b28346072 Download complete
f54e77052ec4 Download complete
c6a5d0807311 Download complete
[*] Building 5.4s (8/8) FINISHED

[+] Building 5.4s (8/8) FINISHED                                docker:desktop-linux
-> transferring dockerfile: 100B
-> [airflow-init internal] load metadata for docker.io/apache/airflow:2.7.0
-> [airflow-init] authn: docker.io/pull token for registry-1.docker.io
-> [airflow-init internal] load .dockerignore
-> transferring context: 3B
-> [airflow-init 1/2] FROM docker.io/apache/airflow:2.7.0@sha256:7a1d5453d511e6d20093c34dbaa0417a3218a7162d01d4c51056d0d201d4e
-> resolve docker.io/apache/airflow:2.7.0@sha256:7a1d5453d511e6d20093c34dbaa0417a3218a7162d01d4c51056d0d201d4e
-> CACHED [airflow-init 2/2] RUN pip install --no-cache-dir airflow
-> [airflow-init] exporting to image
-> exporting layers
-> exporting manifest sha256:161a3c1297a9e1d06771a0d0c93107c180e072e180a9ffc8201400991770d3
-> exporting config sha256:4c7179567f89916d959d21091a0b15493d3d0e0d74302593594e415d080da
-> exporting attestation manifest sha256:187938a70e2777c121f936c94e912135a0a01577cc1cc44939dcf30ba3
-> exporting manifest list sha256:7993b0c1f1ef18400000a34487004f7240bc47784393d4d1009212b0eac1a
-> naming to docker.io/library/airflow-airflow-init:latest
-> unpacking to docker.io/library/airflow-airflow-init:latest
-> [airflow-init] resolving provenance for metadata file
```

Для развертывания nifi используем команду:

\$ docker compose -f docker-compose.nifi.yaml up --build -d

```
\vscode\labs\prerequisites> docker compose -f docker-compose.nifi.yaml up --build -d
apache-nifi Pulled
 444d70ee54 Already exists
 76dd98e6ce9d Download complete
 42969f081378 Download complete
 400f98dc5a1f Download complete
 ec22c511c853 Download complete
 880e41c32028 Download complete
 c861ad0b284d Download complete
 f52c88e7721e Download complete
 990e912b04af Download complete
 b77878610a21 Download complete
 da0d0f13975a Download complete
 364128961ff8 Download complete
Container nifi-apache-nifi-1 Started
PS C:\Users\ikasp\vscode\labs\prerequisites>
```

Для развертывания elasticsearch используем команду:

\$ docker compose -f docker-compose.elasticsearch.yaml up --build -d

```
[+] Running 7/7 sp\vscode\labs\prerequisites> docker compose -f docker-compose.elasticsearch.yaml up --build -d
elasticsearch-kibana Pulled
 12cca292b13c Download complete
 a2f2f93da482 Download complete
 4ef79ea5ec0f Download complete
 1efc276f4ff9 Download complete
 e0d30173d675 Download complete
 d73cf48caaac Download complete
[+] Running 1/1
Container elasticsearch-elasticsearch-kibana-1 Started
PS C:\Users\ikasp\vscode\labs\prerequisites>
```

Для развертывания postgresql используем команду:

\$ docker compose -f docker-compose.postgresql.yaml up --build -d

```
[+] Running 29/29 \vscode\labs\prerequisites> docker compose -f docker-compose.postgresql.yaml up --build -d
adminer Pulled
 b4d1ff79abf6 Download complete
 74ce29a48801 Download complete
 987c568a0d55 Download complete
 51aa935720cd Download complete
 7264a8db6415 Download complete
 a5a6a9033418 Download complete
 33ba53f44015 Download complete
 e3417fc82c92 Download complete
 484c9fcfb07b Download complete
 6091856a6145 Download complete
 85e541ea234a Download complete
 4c1daaf964a4 Download complete
 30b54598555e Download complete
 b0bd67b7da2c Download complete
postgresql-standalone Pulled
 8d0105edc189 Download complete
 dd9adaf025af Download complete
 c9b7bebf3fc9 Download complete
 7f0ae1058953 Download complete
 42f7f2f8d000 Download complete
 a1a96be070c0 Download complete
 ec308293f07d Download complete
 586c52e1f86d Download complete
 55d33c760d25 Download complete
 8a02bd7f95ec Download complete
 da899a2e2339 Download complete
 2628a0547b3c Download complete
 2895a894691d Download complete
[+] Running 3/3
Volume "postgresql_postgres-db-workspace-volume" Created
Container postgresql-postgresql-standalone-1 Started
Container postgresql-adminer-1 Started
PS C:\Users\ikasp\vscode\labs\prerequisites>
```

Для развертывания mlflow используем команду:

\$ docker compose -f docker-compose.mlflow.yaml up --build -d

```
PS C:\Users\ikasp\vscode\labs\prerequisites> docker compose -f docker-compose.mlflow.yaml up --build -d
time="2024-12-03T01:31:08+04:00" level=warning msg="C:\Users\ikasp\vscode\labs\prerequisites\docker-compose.mlflow.yaml: the attribute 'version' is obsolete, it will be ignored, please remove it"
[+] Running 23/23
   db Pulled
   ├── 771e5f80ccc26 Download complete
   ├── a316717fc6ee Download complete
   ├── a1f742e3aa43 Download complete
   ├── b64762744f75 Download complete
   minio Pulled
   ├── 5536c0b07d06 Download complete
   ├── 3440aa9567dd Download complete
   ├── f2f8f30a646a Download complete
   ├── c1d0e26236f5 Download complete
   ├── 4414594dd510 Download complete
   ├── c1cc85e2da65 Download complete
   ├── d57a4fe62ee8 Download complete
   ├── 2b027acd57fe Download complete
   ├── 48e0cfc0f68 Download complete
   mc Pulled
   ├── 3948c4ab0580 Download complete
   ├── 6ab3d7bcedb9 Download complete
   ├── 3398cc2759de Download complete
   ├── fcc8361add52 Download complete
   ├── 2e88e545c999 Download complete
   ├── 8db92fa643ea Download complete
   ├── cc52df538a25 Download complete
[+] Building 45.3s (7/7) FINISHED
=> [web internal] load build definition from Dockerfile
=> => transferring dockerfile: 122B
=> [web internal] load metadata for ghcr.io/mlflow/mlflow:v2.7.1
=> [web internal] load .dockerignore
=> => transferring context: 2B
=> [web 1/2] FROM ghcr.io/mlflow/mlflow:v2.7.1@sha256:6a44110dbb24042577031d14aa1c22e17f82005789640ed7ed326157669272ec
=> => resolve ghcr.io/mlflow/mlflow:v2.7.1@sha256:6a44110dbb24042577031d14aa1c22e17f82005789640ed7ed326157669272ec
=> => sha256:d0c979c5c0baa204a0b07ad3fc7876e72eded7bdb69c8e2034ec7fe000b7021f 245B / 245B
=> => sha256:a0bfbdbb3c060ece4a41933e53a83105507221232c5d10797cee2a1106f4c57 196.70MB / 196.70MB
=> => sha256:6c3f3a0dca6bd365dc1e173443e4851dc9cd2560629222af050058c04bc01af 3.37MB / 3.37MB
=> => sha256:d0bca8f04bae51e1505d3bdabe73302c3e24a1610c1894afcaa643391fb71330 11.54MB / 11.54MB
=> => sha256:b48a735bdcaff937418ddb0b974c9a0f8f6e2a6c26af21a82e328d4ed7cd7 1.08MB / 1.08MB
=> => sha256:7d97e254a0461b0a30b3f443f1daa0d620a3cc6ff4e2714cc1cfd96ace5b7a7e 31.42MB / 31.42MB
=> => extracting sha256:7d97e254a0461b0a30b3f443f1daa0d620a3cc6ff4e2714cc1cfd96ace5b7a7e
=> => extracting sha256:b48a735bdcaff937418ddb0b974c9a0f8f6e2a6c26af21a82e328d4ed7cd7
=> => extracting sha256:d0bca8f04bae51e1505d3bdabe73302c3e24a1610c1894afcaa643391fb71330
=> => extracting sha256:d0c979c5c0baa204a0b07ad3fc7876e72eded7bdb69c8e2034ec7fe000b7021f
=> => extracting sha256:6c3f3a0dca6bd365dc1e173443e4851dc9cd2560629222af050058c04bc01af
=> => extracting sha256:a0bfbdbb3c060ece4a41933e53a83105507221232c5d10797cee2a1106f4c57
=> [web 2/2] RUN python -m pip install --no-cache-dir pymysql
=> [web] exporting to image
=> => exporting layers
=> => exporting manifest sha256:d8c55a291cea3ee86fb03391df1b658a513af0013be9ac5423c3e3675d677094
=> => exporting config sha256:71328dca9ba4cdadd5403afe15af5ae7698669b0802363c0bed62d3fdd5c78
=> => exporting attestation manifest sha256:43bd63dbaa078016a4adfd398058f065453190021c142daa1dffe61b7829a3
=> => exporting manifest list sha256:24f93e30feefee83e9c3cae3df8b3bf33c00f5662eb04722b05d0852894b86fe
=> => naming to docker.io/library/mlflow-web:latest
=> => unpacking to docker.io/library/mlflow-web:latest
=> [web] resolving provenance for metadata file
[+] Running 4/4
   Volume "mlflow_minio_data" Created
   Volume "mlflow_dbdata" Created
   Container mlflow-db-1 Started
   Container mlflow-minio-1 Started
   Container mlflow-mc-1 Started
   Container mlflow_server Started
PS C:\Users\ikasp\vscode\labs\prerequisites>
```

Посмотрим в докер, и убедимся, что все запущено и работает.

Containers [Give feedback](#)

Container CPU usage

10.18% / 800% (8 CPUs available)

Container memory usage

5.38GB / 7.53GB

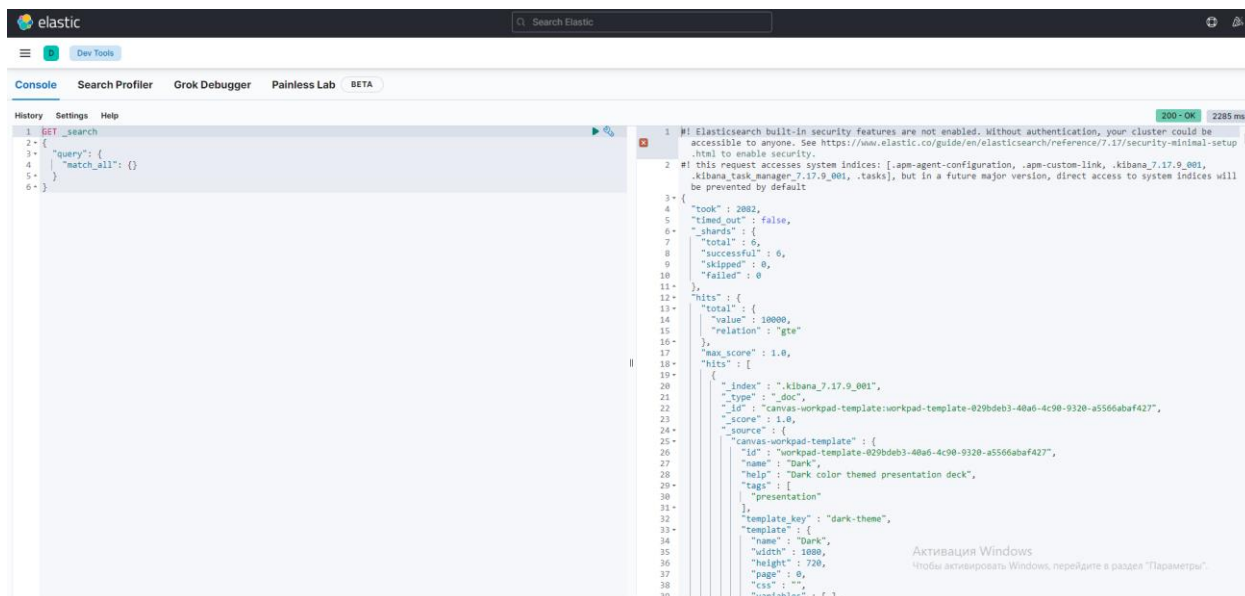
Show charts

Q Search

Only show running containers

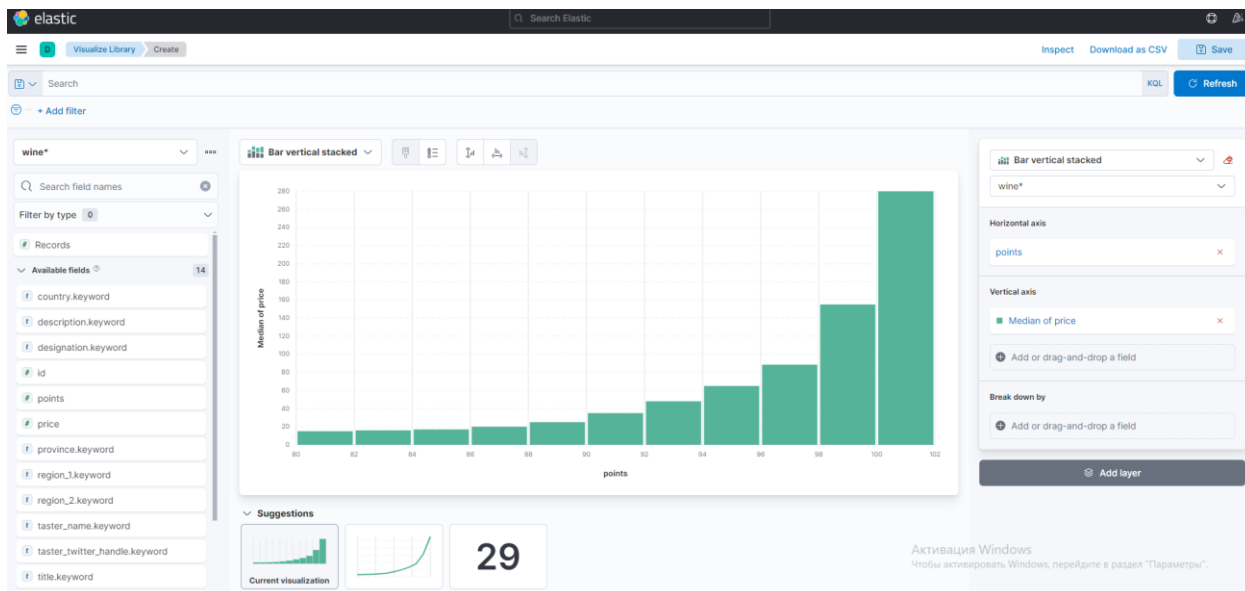
<input type="checkbox"/>	Name	Container ID	Image	Port(s)	CPU (%)	Last started	Actions
<input type="checkbox"/>	airflow	-	-	-	8.31%	3 hours ago	<div><div></div><div></div><div></div></div>
<input type="checkbox"/>	redis-1	333868496cb0	redis:latest		0.29%	4 hours ago	<div><div></div><div></div><div></div></div>
<input type="checkbox"/>	postgres-1	0ba5656871d1	postgres:13		1.88%	4 hours ago	<div><div></div><div></div><div></div></div>
<input type="checkbox"/>	docker-proxy-1	5d564c122e58	docker:24-dind		0%	3 hours ago	<div><div></div><div></div><div></div></div>
<input type="checkbox"/>	airflow-init-1	79c6130eae89	airflow:airflow-init		0%	3 hours ago	<div><div></div><div></div><div></div></div>
<input type="checkbox"/>	airflow-webserver-1	397d1ef0ce40	airflow:airflow-webserver	8080:8080	0.25%	3 hours ago	<div><div></div><div></div><div></div></div>
<input type="checkbox"/>	airflow-triggerer-1	75786c777e51	airflow:airflow-triggerer		0.8%	3 hours ago	<div><div></div><div></div><div></div></div>
<input type="checkbox"/>	airflow-worker-1	66eba46ea575	airflow:airflow-worker		0.2%	3 hours ago	<div><div></div><div></div><div></div></div>
<input type="checkbox"/>	airflow-scheduler-1	1aa97c1fc9c9	airflow:airflow-scheduler		4.89%	3 hours ago	<div><div></div><div></div><div></div></div>
<input type="checkbox"/>	nifi	-	-	-	0.55%	3 hours ago	<div><div></div><div></div><div></div></div>
<input type="checkbox"/>	elasticsearch	-	-	-	1.25%	3 hours ago	<div><div></div><div></div><div></div></div>
<input type="checkbox"/>	postgresql	-	-	-	0.02%	7 minutes ago	<div><div></div><div></div><div></div></div>
<input type="checkbox"/>	mlflow	-	-	-	0.04%	2 minutes ago	<div><div></div><div></div><div></div></div>
<input type="checkbox"/>	db-1	eccca770579d	mysql/mysql-server:5.7.28	3306:3306	0.03%	2 minutes ago	<div><div></div><div></div><div></div></div>
<input type="checkbox"/>	minio-1	8b674b3c3d31	minio/minio	9000:9000	0%	2 minutes ago	<div><div></div><div></div><div></div></div>

По окончании работы DAG-а, вытягиваем все данные в Elasticsearch.



The screenshot shows the Elasticsearch console interface. On the left, a search query is entered: `{ "query": { "match_all": {} } }`. On the right, the search results are displayed in JSON format. The results include a `took` value of 2802, a `timed_out` value of false, and a `total` of 6. The `hits` array contains one document with the following fields: `_index`: ".kibana_7.17.9_001", `_type`: ".doc", `_id`: "canvas-workpad-template:workpad-template-029deb3-40a6-4c90-9320-a5566abaf427", `_score`: 1.0, and `_source`: { "canvas-workpad-template": { "id": "workpad-template-029deb3-40a6-4c90-9320-a5566abaf427", "name": "Dark", "help": "Dark color themed presentation deck", "tags": ["presentation"], "template_key": "dark-theme", "template": { "name": "Dark", "width": 1080, "height": 720, "page": 0, "css": "", "variables": { "f": 1.0. The console also shows a warning message about Elasticsearch security features.

Построение графика зависимости цены вина от количества его очков.

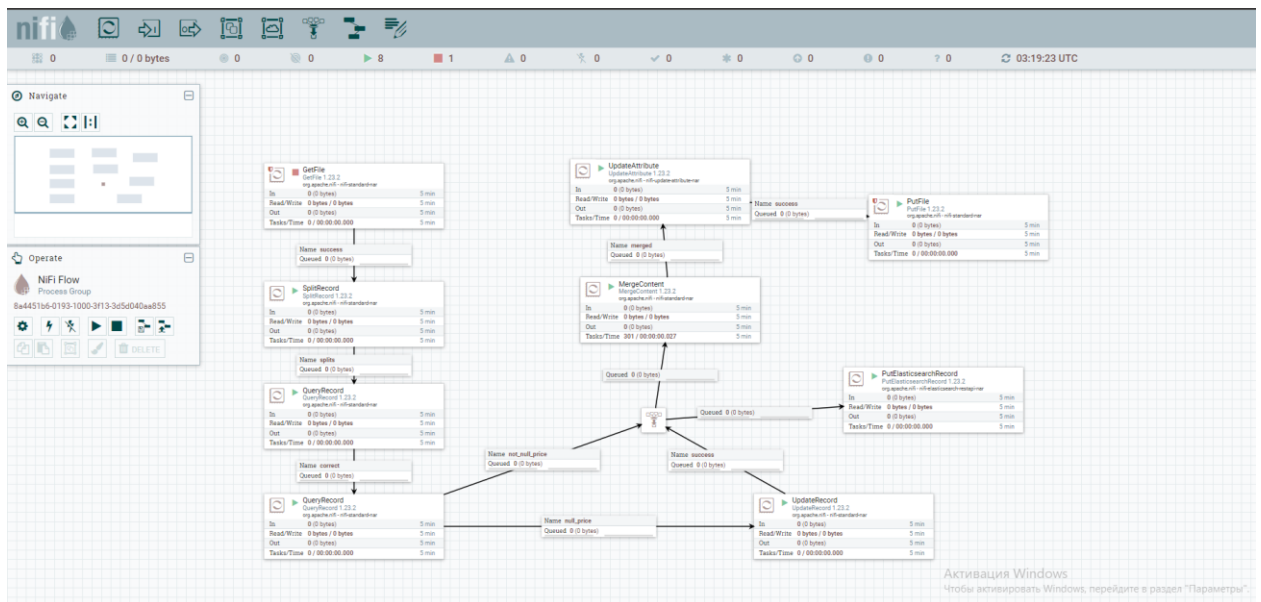


2. Apache Nifi

Для реализации пайплайна использовались следующие процессоры:

- GetFile
- SplitRecord
- QueryRecord
- UpdateRecord
- MergeContent
- PutFile
- PutElasticsearchRecord

Конечный вид пайплайна:



Заключение

Данная лабораторная работа вызвала кучу затруднений. Всего у меня заняло 4 дня на реализацию. У меня не было опыта работы с данными инструментами и пришлось очень долго вникать и разбираться, методом проб и миллионами ошибок. Много ошибок было при запуске Airflow, от “несуществования” директории до “Connection error”. Еще сложнее было работать с nifi. Очень сложный интерфейс процессоров, лично я не понимал, когда процессор завершил свою работу. Ориентироваться по Tasks/Time невозможно, значения постоянно меняются, то в большую, то в меньшую сторону.