Apache Airflow

1. Написание кода DAG’a.

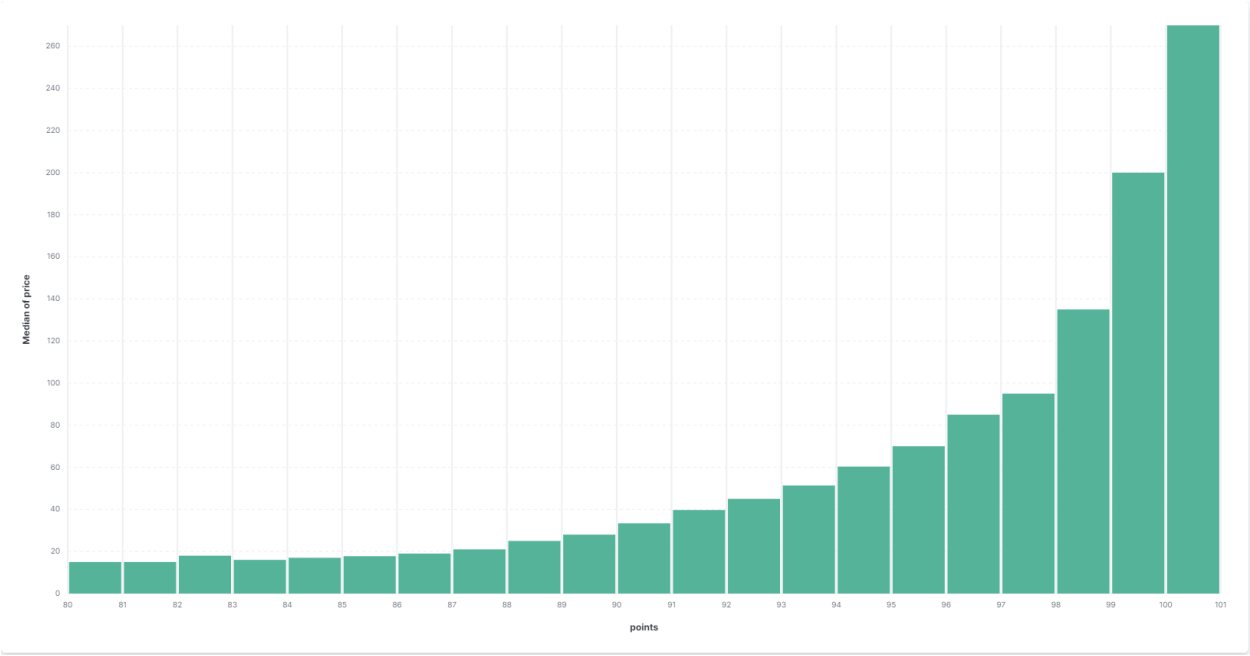
Были написаны python функции, которые были потом привязаны к таскам.

В целом проблем не было за исключением, что было изначально непонятно как передавать данные между тасками и проблемы с kibana. В первой реализации я попробовал использовать xcom, но иногда вылетали ошибку из-за недостаточного количества памяти, поэтому я переделал, чтобы промежуточные данные сохранялись .csv файлы в папке data, и следующие таски брали эти данные из этих .csv файлов. Проблема с kibana заключалась в том, что данные не сохранялись, решилась она предварительным переформатированием строки в json.

2. Apache nifi.

Мягко говоря, apache nifi мне совсем не понравился в связи с огромным количеством проблем. Одна из них – это крайне скудная документация к процессорам, да и в целом информации о данном инструмента довольно мало. По ходу выполнения были выявлены неочевидные особенности, такие как: невозможность работы с большим количеством файлов в рамках одного процессора, приходилось заранее объединять их в подгруппы, невозможность объединить несколько файлов с разными названиями, необходимо предварительно переименовать все файлы на одно имя.

Используя данные, обработанные с помощью пайплайна, в была построена данная гистограмма.



График