РАЗРАБОТКА СЕРВИСА ОБРАБОТКИ ПОИСКОВЫХ ЗАПРОСОВ ОНЛАЙН БРОНИРОВАНИЙ ОТЕЛЕЙ ДЛЯ ОЦЕНКИ АНАЛИТИЧЕСКИХ ПОКАЗАТЕЛЕЙ

DEVELOPMENT OF A SERVICE FOR PROCESSING SEARCH QUERIES FOR ONLINE HOTEL BOOKINGS FOR EVALUATING ANALYTICAL INDICATORS

УДК 004

Яблонский Дмитрий Анатольевич, бакалавр, ФГБОУ ВО МГТУ «СТАНКИН»

Dmitry Yablonsky, bachelor, FGBOU VO MSTU «STANKIN», dmitryaposh@yandex.ru

Аннотация

Статья посвящена описанию разработанного программного обеспечения, направленного на получение актуальных данных по бронированиям отелей для определения тенденций, прогнозирования производственных показателей и оптимизации собственных расходов. Поставлена актуальность проблемы обработки больших данных. Описаны требования к задаче создания программы. Представлена схема системы, разобраны основные ее составляющие и примерный алгоритм работы.

Annotation

The article is devoted to the description of the developed software aimed at obtaining up-to-date data on hotel reservations for determining trends, forecasting production indicators and optimizing their own costs. The urgency of the problem of processing big data is stated. The requirements for the task of creating a program are

described. The diagram of the system is presented, its main components and an approximate algorithm of operation are analyzed.

Ключевые слова: программное обеспечение, обработка данных, поисковые запросы, аналитические показатели.

Key words: software, data processing, search queries, analytical indicators.

Что такое туризм с большими данными?

Индустрия туризма процветает на информации. Большие данные могут предоставлять актуальные и глубоко обоснованные выводы о поведении и человеческой деятельности, что способствует развитию индустрии туризма. Туристы оставляют после себя различные цифровые следы при использовании мобильных технологий в сети. Каждый турист получает огромное количество данных обо всем, что имеет отношение к различным этапам путешествия - до, между и после путешествия. Большая часть информации носит внешний характер, например, в социальных сетях или в форме Twitter.

Из-за наличия огромных объемов данных в облаке аналитика стала необходимостью для понимания информации, содержащейся в данных. Если вы новый клиент, планирующий поездку, вы, вероятно, приобретете больше источников в Интернете, когда покупаете билеты, исследуете достопримечательности или бронируете жилье. Представители туристической индустрии постепенно обращаются к большим данным, чтобы найти новые способы улучшения возможностей, принятия решений и общей производительности, например взаимосвязь разрозненной информации может быть сделана возможной с помощью больших данных.

Продавцы и агентства, занимающиеся туризмом, могут найти различные методы использования разнообразных ресурсов данных для взаимодействия с потенциальными гостями на каждом этапе тура и использовать эти большие источники данных для своевременного и лучшего понимания растущей статистики посетителей. Они также могут проводить перепродажу, чтобы

нацелить клиентов, выбравших конкретный пункт назначения на веб-сайте туристического агентства.

Новые информационные потоки

Большие данные о туризме — это типичная информация, созданная самими путешественниками. Он улучшает мыслительный процесс туристического бизнеса, анализируя спрос клиентов на различные услуги и продукты.

Надежность

Большие данные основаны не на опросах, а на реальных действиях пользователей. Другими словами, эти действия - а не заявленные запросы или намерения - были изучены. Собирая все источники данных вместе, большие данные на несколько порядков расширяют базу выборки, на которой обычно создаются традиционные основы.

Все мы живем в период беспрецедентного изменения убеждений клиентов, бизнес-моделей компаний и реакции потребителей, вызванных технологиями, которые одновременно разрушают существующие институты и создают новые. В любом случае, большие данные о туризме указывают на заметные изменения во взаимоотношениях между клиентами и бизнесом. Мы можем использовать большие данные для поддержки опыта и предоставления превосходной прогнозной аналитики для улучшения решений и ожиданий клиентов.

Описание разработанной системы

С целью повышения качества выявления ценной информации из большого количества структурированных наборов сообщений в туристической сфере был разработан сервис обработки поисковых запросов онлайн. При его разработке были поставлены следующие требования:

1. Получение метапоисковых запросов в формате json.
2. Группировка и агрегация полученных запросов.
3. Получение цен конкурентов, по наиболее интересным для нас запросам.
4. Формирование общей структуры и отправка в базу данных.

Для разработки сервиса Scheduler был использован язык программирования scala, распределённый программный брокер сообщений Apache kafka, Flink - фреймворк для масштабируемой распределенной обработки потоков.

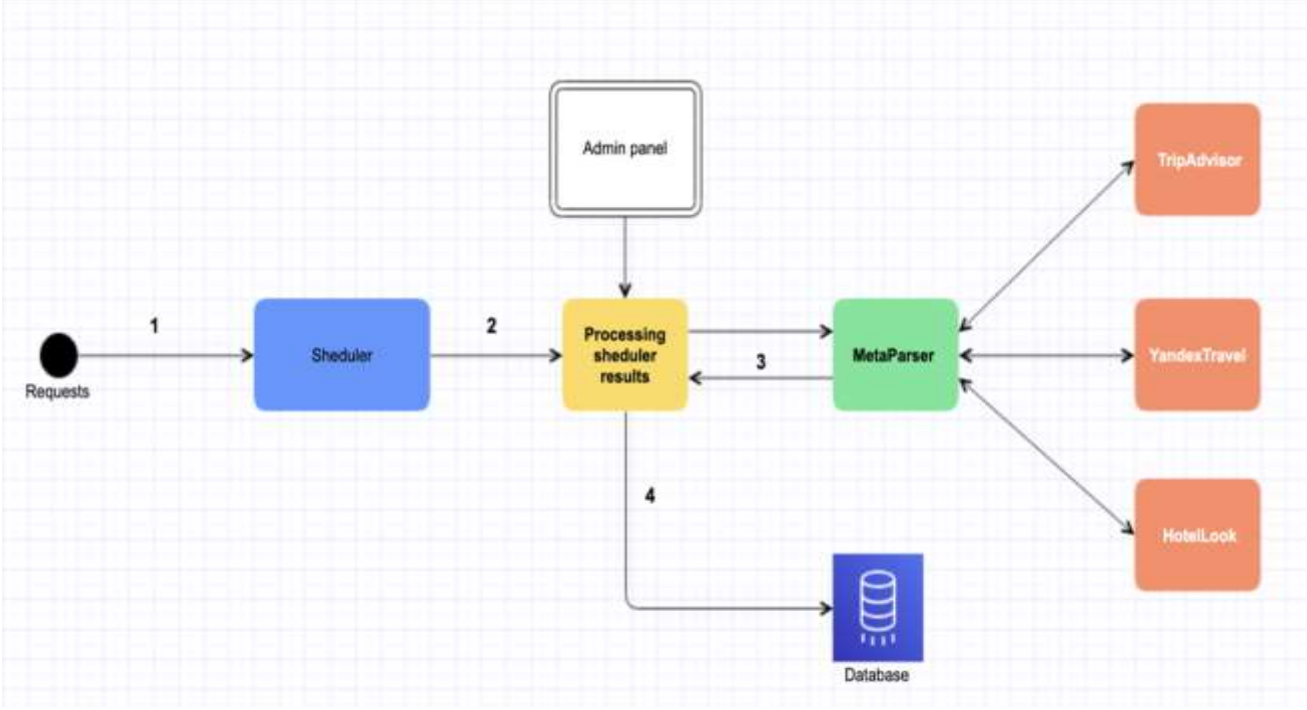


Рис. 1 – Схема процесса

Выше представлена схема процесса (рис. 1), который предоставит аналитические показатели по наиболее запрашиваемым параметрам отелей. Составляющие процесса:

- Requests – kafka топик с запросами (1), которые привели на наш ресурс с метапоисковиков, таких как TripAdvisor.
- Scheduler – сервис, производящий обработку больших данных и генерирующий задания на обработку наиболее интересной для нас информации. Задания также передаются через кафка топик (2)
- Processing scheduler results – сервис, который выстраивает очередь с приоритетом из тех заданий который получает от scheduler и передает запрос в MetaParser. С помощью http запроса (3)

- MetaParser – спрашивает на различных платформах доступность по необходимым отелям, агрегирует всю информацию и отправляет назад в Processing scheduler results.
- Database – база данных для сохранения аналитических показателей.

Нас больше всего интересует сервис scheduler, т. к. именно он производит обработку больших данных. Ключевые моменты заключаются в постоянных группировках запросов в 1-минутные и 5-минутные окна «DataStream», внутри которых происходит агрегация. 5-минутные используются для отправки данных в InfluxDB (база данных временных рядов) для обработки запросов и построения метрик. Как сказано в требованиях, запросы из requests поступают в формате json и в 1-минутном окне происходит их группировка и фильтрация по различному набору параметров (рынок, отель, город, страна, партнер).

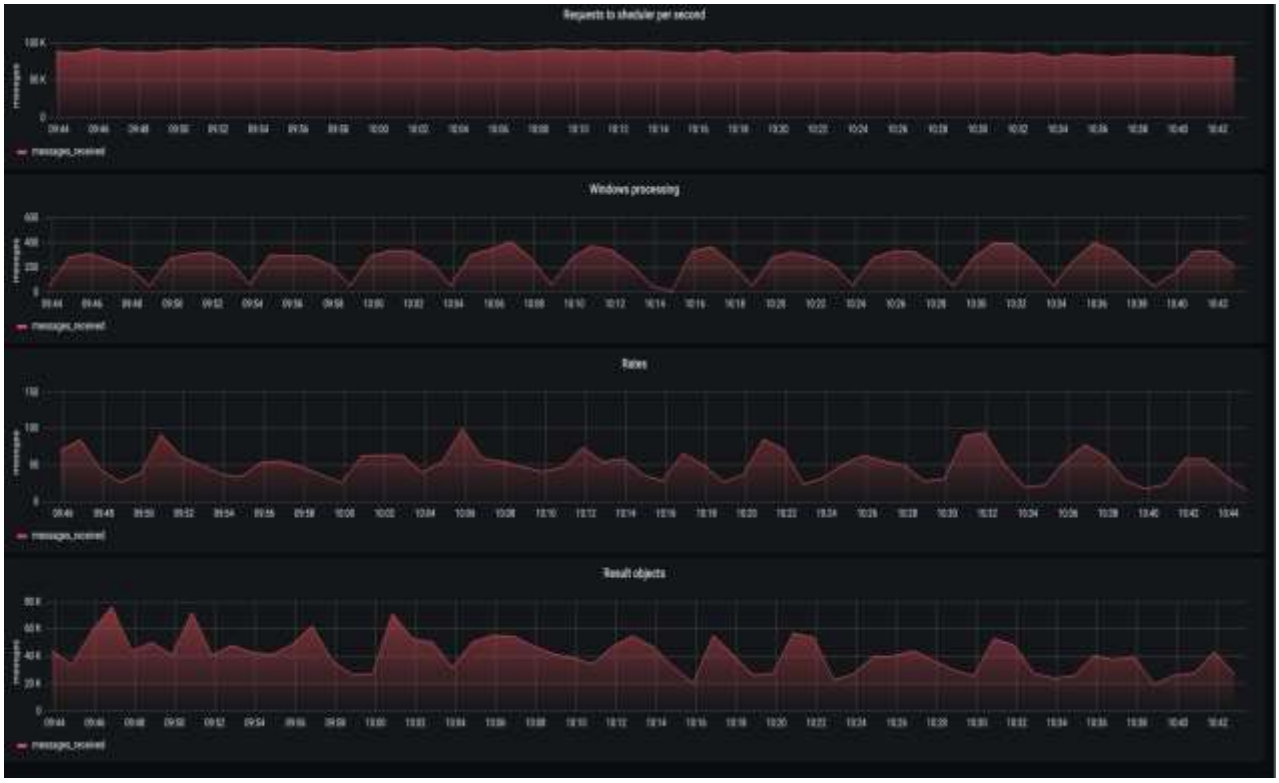


Рис. 2 – технические показатели сервиса.

После настройки и отладки метрик можно получать достаточно подробную информацию о работе сервиса. Например, как на рисунке 2.

Результат всего процесса можно выстроить по-разному. В таблице 1 представлен ряд параметров, которые можно получить посредством выгрузки из базы данных. (это не полный список, но наиболее интересные для прогнозирования производственных показателей)

timestamp	когда была создана задача на парсинг отеля на метапоиске. Т. е. это время за несколько минут до парсинга (UTC)
checkin	дата заезда в отель, указанная на метапоиске
checkout	дата выезда из отеля, указанная на метапоиске
metasearch	метапоиск на котором парсился отель
market	из какого региона делался запрос на метапоиск
adults	число взрослых в запросе на метапоиске
currency	валюта, выбранная на метапоиске
hotel_pos	позиция отеля в поисковой выдаче на метапоиске
city	город отеля
country	страна отеля
metasearch_hotel_id	идентификатор отеля на метапоиске
metasearch_hotel_name	название отеля на метапоиске
ourhotel_id	наш идентификатор отеля
service_agent	название нашего агента, продающего отель
other_agent	название лучшего конкурента на метапоиске
our_price	наша цена
other_price	цена лучшего конкурента
date	дата записи данных в кафку (UTC)
hour	час записи данных в кафку (UTC)

Таб. 1 – результирующие параметры

Заключение

Благодаря такому подходу к обработке больших данных можно определять тенденции, прогнозировать производственные показатели и оптимизировать собственные расходы в различных бизнес-процессах.

Литература

1. Envisioning Tourist Demand with Big Data
2. Flink - https://flink.apache.org/
3. Influx with grafana - https://grafana.com/docs/grafana/latest/getting-started/getting-started-influxdb/

Literature

1. Envisioning Tourist Demand with Big Data
2. Flink - https://flink.apache.org/
3. Influx with grafana - https://grafana.com/docs/grafana/latest/getting-started/getting-started-influxdb/