Тестовое задание на позицию дата-аналитика

Вводные по тестовому заданию

По ссылке (https://disk.yandex.ru/d/gmqRNztpfIUEww) вы можете скачать gzip-архив, в котором находится текстовый лог-файл с HTTP-запросами к веб-серверу.

Формат файла:

- одна строка – один HTTP-запрос,
- поля в строке разделены символом табуляции.

Описание полей:

- **datetime** – дата и время запроса (GMT+0),
- **host** – наименование домена (сайта), к которому был совершен запрос,
- **remote_addr** – IP-адрес источника запроса,
- **geo_country_code_variable** – двухбуквенный код страны источника запроса,
- **connection_requests** – порядковый номер запроса в рамках установленного соединения,
- **connection** – порядковый номер соединения,
- **ssl_protocol** – используемый SSL/TLS -протокол,
- **request_time** – время в секундах, потраченное на обработку запроса,
- **body_bytes_sent** – количество переданных данных в рамках запроса (байты),
- **request_method** – HTTP-метод запроса (обычно GET, POST, но могут быть и другие),
- **request_uri** – первоначальная строка запроса целиком (путь),

- **server_port** – TCP-порт, на который пришел запрос (80 или 443),
- **http_referer** – значение заголовка Referer,
- **http_user_agent** – значение заголовка User-Agent,
- **upstream_cache_status** – статус кэширования (HIT для объекта, отданного из кэша, или MISS для объекта запрошенного с сервера оригинации, другие – редко),
- **status** – HTTP-статус (200 - все хорошо, 3xx - редирект, 4xx - нет доступа, 5xx – что-то не так с сервером),
- **upstream_status** – HTTP-статус, переданный сервером выше,
- **upstream_response_time** – время (секунды), потраченное на обработку запроса на сервере выше,
- **upstream_addr_list** – список вышестоящих серверов (может быть пустым, если объект отдан из кэша – HIT).

Задание

1. Визуализация

1. Постройте общий график запросов: ось X - время, ось Y - количество запросов.
2. Постройте графики количества запросов в зависимости от их группы. Группа запросов определяется первой цифрой: 2xx, 3xx, 4xx, 5xx. Можно на одном графике. Ось X - время, ось Y - количество запросов определенной группы.
3. Постройте график среднего времени ответа. Значение нужно брать из поля **request_time**, агрегировать запросы нужно за 5-ти минутный период. Ось X - время, ось Y - среднее время ответа в секундах.

2. Анализ

1. Определите, были ли аномалии в количестве запросов и временной промежуток, когда они наблюдались.

2. Определите паттерны аномалий: характерные признаки запросов, которые относят их к аномалии.
3. Опишите ход ваших рассуждений по определению аномалий и их признаков.
4. Опишите алгоритм, который позволит в будущем исключать аномальные запросы такого характера без ручного анализа лог-файла.

Представление результата

Результат пришлите в pdf-файле с необходимыми графическими и текстовыми.