Тестовое задание на позицию Data Engineer, Одноклассники

В приложенном файле feeds\_show.json.gz вы найдете пример данных по показам в ленте в сети Одноклассники. Одна строчка – показ одного «фида» в ленте. Фидом называем один логический элемент (пост друга, информация о загрузке фото и т.д.)

Комментарии по некоторым полям:

* durationMs – время в течение которого пользователь видел фид
* position – позиция в ленте на которой был показан фид
* owners – создатели контента в фиде
* resources – ссылки на ID контента, содержащегося в фиде

**Задание**

1. Опишите схему данных в любой удобной вам нотации.
2. Предложите набор тестов для входных данных, которые максимально гарантировали бы их валидность. Желательно приложить какой-либо прототип их реализации, но подойдет и простое описание логики их работы.
3. На основании этих данных мы хотим считать следующий набор метрик. Напишите алгоритмы / запросы для их расчета. Предпочтительно на SQL / Spark DataFrame API. Реальный объем данных – миллиарды строк в день, это необходимо учитывать при написании запросов.

Если это необходимо и, на ваш взгляд, является оптимальным – допускается создание промежуточных таблиц или других необходимых объектов.

Можно исходить из предположения что данные уже переложены из json в какой-то более удобный формат / таблицу и партиционированы по дням.

* 1. Количество показов и уникальных пользователей за день в разрезе по платформам, в том числе по всем платформам суммарно;
  2. Количество за день уникальных авторов и уникального контента, показанного в ленте;
  3. Количество сессий, средняя глубина просмотра (по позиции фида) и средняя продолжительность пользовательской сессии в ленте за день. Определение сессии остаётся на ваше усмотрение;
  4. Дополнительно дан список пользователей в формате <userId: Long>, включающий в себе несколько тысяч пользователей. Нужно посчитать количество показов фидов в ленте по этим пользователям.

1. Оцените сложность алгоритмов из пункта 3 в терминах О-большое. Можно ли уменьшить их сложность, изменив логику / формат логирования данных?