Добрый день! Перед Вами тестовое задание на должность разработчика BI компании СКБ Контур. Здесь собраны задачи разного уровня сложности: от элементарных до тех, что требуют глубоких знаний архитектуры SQL Server. Поэтому если на какой-либо из вопросов не найдется ответа, то можно честно в этом признаться и решать следующие задания. Ваша задача – показать общий уровень владения предметной областью. Итак, поехали…

**Введение**

Представим, что в конце 2018 года после строгого отбора Вы были приняты в штат элитной авиакомпании AdventureBirds на должность эксперта по базам данных. От рядовых разработчиков Вы получили имеющуюся документацию на систему из которой узнали следующее…

**Описание БД**

Система развернута на SQL Server 2017, Enterprise Edition. Реляционная база данных находится в полной модели восстановления. Схема данных представлена на рис.1.



**Рис.1. Структура базы данных.**

Flights – основная и единственная таблица фактов. Все остальные таблицы являются справочниками. Описание полей таблицы Flights представлено в таблице 1.

**Таблица 1. Описание полей таблицы Flights.**

|  |  |  |
| --- | --- | --- |
| **№ п/п** | **Наименование поля** | **Описание поля** |
| 1 | RecordId | Идентификатор строки – искусственный, нарастающий идентификатор |
| 2 | DepartureDateTime | Дата/время отправления |
| 3 | ArrivalDateTime | Дата/время прибытия |
| 4 | DepartureAirportId | Аэропорт отправления |
| 5 | DestinationAirportId | Аэропорт назначения |
| 6 | AirCompanyId | Идентификатор авиакомпании |
| 7 | PlaneId | Идентификатор самолета |
| 8 | Passengers | Пассажиров на рейсе |

В целях логистического анализа в таблице хранятся данные по всем авиакомпаниям. Ваша авиакомпания представлена в таблице под идентификатором 0. Данные в системе хранятся с 2010 года, общий объем данных составляет порядка 2 млрд.стр., занимаемое место на дисках – около 700 Гб.

**Описание куба**

Увы, по какой-то неизвестной причине подробное описание куба было утеряно. Единственное, что удалось выяснить – куб создавался для расчета пассажиропотока (суммарного количества перевезенных пассажиров) в разрезе дат, авиакомпаний и типов ВС (самолетов).

**Загрузка данных**

Данные в реляционную базу загружаются каждую ночь SSIS-пакетом. Данные в таблицах не перезаписываются и не удаляются, загружаются только новые записи. Суточный прирост данных – порядка 4 млн.стр., данные загружаются в таблицу примерно в течение 40 мин.

После наполнения реляционной БД следующим шагом задания запускается обработка куба. Измерения обрабатываются полностью, группы мер – только добавляются новые данные. Записи для куба выбираются напрямую из основных таблиц реляционной БД. Обработка куба занимает порядка 5 мин.

**Текущая нагрузка на систему**

В настоящее время пользователи обращаются к данным только через куб для расчета пассажиропотока в различных разрезах и за различные периоды.

**Планы развития системы**

В ближайшее время планируются следующие задачи:

1. Разработка новых регулярных отчетов, с которыми пользователи могли бы работать через web.
   1. Детализация трафика.

Назначение отчета: вывести всю информацию по рейсам за выбранный период и выбранной авиакомпании.

Требования к содержанию отчета: в отчете должны присутствовать поля: «Дата/время отправления», «Дата/время прибытия», «Пункт отправления», «Пункт назначения», «Авиакомпания», «Самолет», «Пассажиров на рейсе», «Загруженность рейса, %» (количество пассажиров / мест в самолете)

Требования к фильтрации данных: для выборки данных должны задаваться три параметра: даты начала / окончания запрашиваемого периода, авиакомпания. Отдел логистики хотел бы, чтобы наименование авиакомпании они выбирали из выпадающего списка. Причем в начале списка должен обязательно присутствовать пункт «Все», который позволил бы выбрать данные по все авиакомпаниям, далее – перечень авиакомпаний в алфавитном порядке. Период запрашиваемых данных не может превышать 1 месяц.

* 1. Логистический отчет.

Назначение отчета: никто не может толком объяснить, как подобная информация будет использоваться, но отдел логистики утверждает, что им обязательно нужен отчет, который позволил бы за текущий месяц выбрать все рейсы, вылетевшие перед рейсами нашей авиакомпании.

Пример исходных данных и результирующей выборки: пример исходной выборки представлен на рис.2. На этих данных мы видим 3 рейса нашей авиакомпании (записи 4, 12, 15). В отчет должны попасть записи 1, 11, 13:

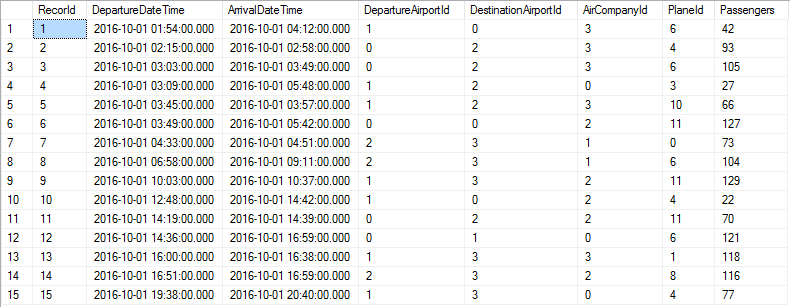
- рейс 4 вылетел из аэропорта 1, перед ним из этого же аэропорта в 01:54 вылетел рейс авиакомпании 3 (запись 1);

- рейс 12 вылетел из аэропорта 0, перед ним из этого же аэропорта в 14:19 вылетел рейс авиакомпании 2 (запись 11);

- рейс 15 вылетел из аэропорта 1, перед ним из этого же аэропорта в 16:00 вылетел рейс авиакомпании 3 (запись 13).

Требования к содержанию отчета: в отчете должны присутствовать поля: «Дата/время отправления», «Дата/время прибытия», «Пункт отправления», «Пункт назначения», «Авиакомпания».

Требования к фильтрации данных: данные фильтруются только по периоду с первого числа месяца по текущее число включительно.



**Рис.2. Пример исходных данных для логистического отчета.**

* 1. Сводный отчет по пассажиропотоку.

Назначение отчета: вывести сводную информацию по пассажиропотоку за прошедший год по всем авиакомпаниям.

Требования к содержанию отчета: в отчете в столбцах должны быть месяца с январь по декабрь прошедшего года. В строках для каждой авиакомпании должен быть посчитан суммарный пассажиропоток для каждого месяца.

Требования к фильтрации данных: данные фильтруются только по периоду с 1-го числа предыдущего месяца - 1 год по заключительное число предыдущего месяца включительно (если, например, отчет формируется 17.05.2019, то период отчета должен быть с 01.05.2018 по 30.04.2019 включительно).

1. Добавление новой меры в куб. Помимо регулярных отчетов отдел логистики также хотел бы добавить в куб новую расчетную меру: средняя загруженность рейсов. Загруженность рейса рассчитывается как отношение перевезенных пассажиров к количеству мест в самолете.
2. В целях высвобождения дисковых ресурсов отдел логистики согласовал удаление старых данных до начала 2017 года. По предварительным оценкам это позволит сократить размер таблицы до 400 млн.стр. и высвободить порядка 0,5 Тб дискового пространства.

**Вопросы**

1. Напишите запрос, с помощью которого Вы бы формировали набор данных для отчета по детализации трафика.
2. Напишите запрос, с помощью которого Вы бы формировали набор данных для логистического отчета.
3. Напишите запрос, с помощью которого Вы бы формировали набор данных для сводного отчета по пассажиропотоку.
4. Как бы Вы реализовывали требования к фильтрации данных для отчета по детализации трафика?
5. Как бы Вы реализовывали сортировку данных в отчетах?
6. К сожалению, документации на куб не нашлось. Но, основываясь на своем опыте, Вы наверняка сможете предположить, какие группы мер и меры, измерения и атрибуты присутствуют в кубе?
7. Напишите MDX или DAX запрос, с помощью которого Вы бы смогли рассчитать новый показатель для куба. Понадобятся ли какие-нибудь дополнительные работы с кубом / базой данных для реализации этого требования?
8. Ознакомившись подробнее со структурой данных, Вы увидели то, отчего волосы на голове зашевелились: основная таблица фактов была без индексов. Какие индексы Вы бы предложили исходя из существующей и планируемой нагрузки на систему? Или, быть может, они тут все-таки не нужны?
9. Вас неуклонно преследует мысль, что загрузка данных SSIS-пакетом выполняется уж очень долго. Какие рекомендации по настройке базы данных / SSIS-пакета Вы могли бы дать разработчикам, чтобы ускорить вставку данных в таблицу Flights.
10. Как бы Вы реализовывали удаление устаревших данных из таблицы Flights.
11. В нескольких предложениях опишите сложившееся впечатление о системе: что в предложенной архитектуре Вам понравилось и что бы Вы в ней поменяли.