Greenplum для разработчиков и архитекторов баз данных

Оптимизация сложных запросов в MPP-кластерах: Greenplum, Arenadata DB, Cloudberry Database

Меня хорошо видно & слышно?

Защита проекта
Тема: Оптимизация сложных запросов в MPP-кластерах: Greenplum, Arenadata DB, Cloudberry Database

Игорь Щербаков

Разработчик баз данных

План защиты

- Цель и задачи проекта
- Какие технологии использовались
- Что получилось
- Выводы
- Вопросы и рекомендации

Цель и задачи проекта

Цель проекта:
Разбор методов оптимизации сложных запросов в MPP-кластерах

1. Собрать методы оптимизации запросов из разных источников
2. Подготовить стенды MPP-кластеров: Greenplum, Arenadata DB, Cloudberry Database
3. Разобрать методы оптимизации на различных примерах
4. Проанализировать планы выполнения запросов

Какие технологии использовались

1. MPP - Massively Parallel Processing
2. Virtualization, Docker
3. Greenplum, Arenadata DB, Cloudberry Database
4. PXF - Platform Extension Framework
5. Dbeaver, PostgreSQL

Использованные MPP-кластеры

Созданные кластеры:

1. Greenplum 6.27 (из исходников)
2. Arenadata DB 7.2 (с помощью ADCM и бандлов)
3. Cloudberry Database 1.6 (из исходников)

Использованные «песочницы»:

1. Greenplum 6.23
2. Cloudberry Database 1.5.1

База данных «Авиаперевозки»

1. Курс “QPT. Оптимизация запросов” фирмы “PostgresPro” (редакция 27.12.2024): https://postgrespro.ru/education/demodb

2. Книга: Домбровская Г., Новиков Б., Бейликова А. Оптимизация запросов в PostgreSQL. - М.: ДМК Пресс, 2022. (PostgreSQL Query Optimization. The Ultimate Guide to Building Efficient Queries. - Apress, 2021.) https://drive.google.com/drive/folders/13F7M80Kf_somnjb-mTYAnh1hW1Y_g4kJ?usp=sharing

8 таблиц, ~22 млн. строк, ~1.5 Gb

ER-диаграмма «Авиаперевозки»

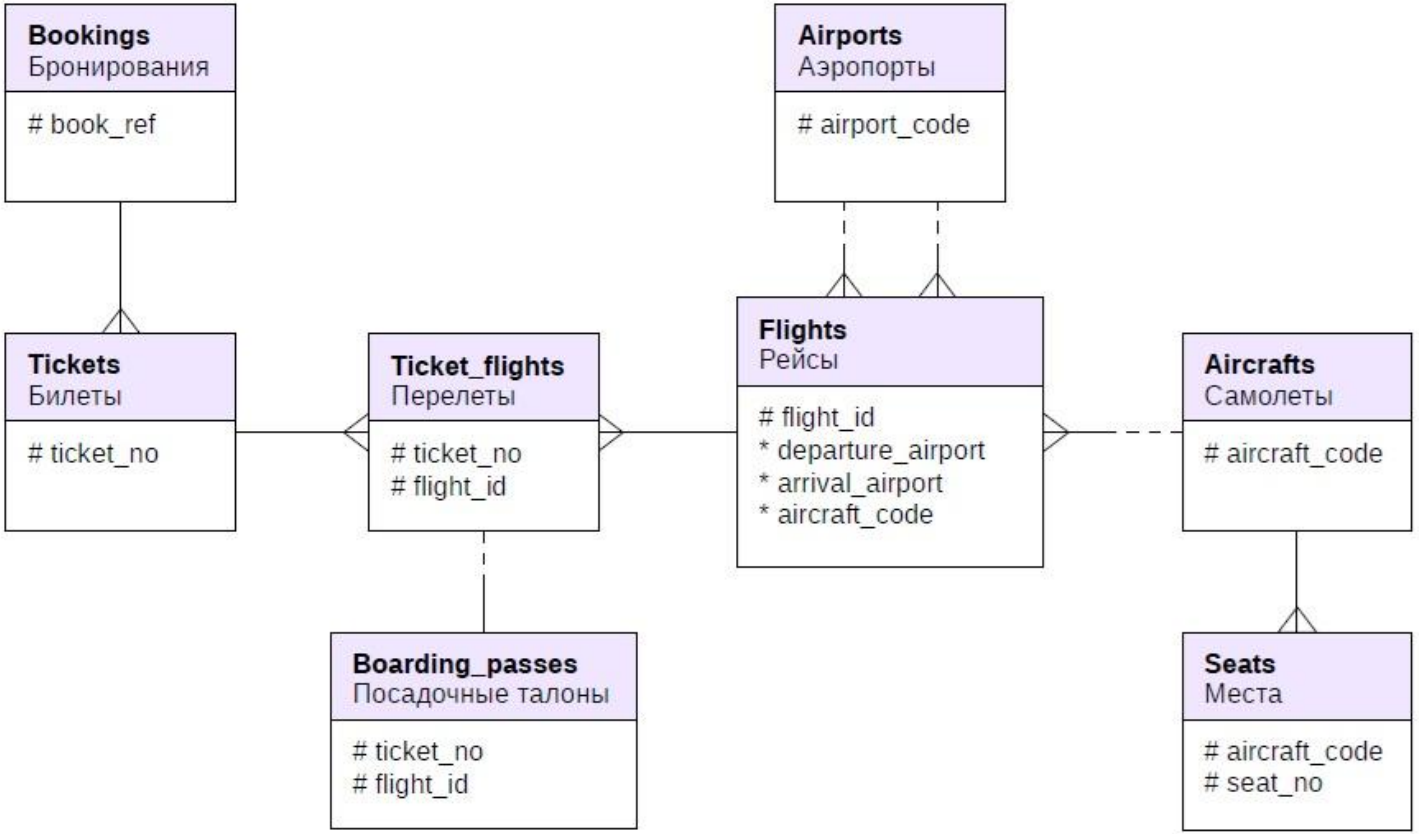
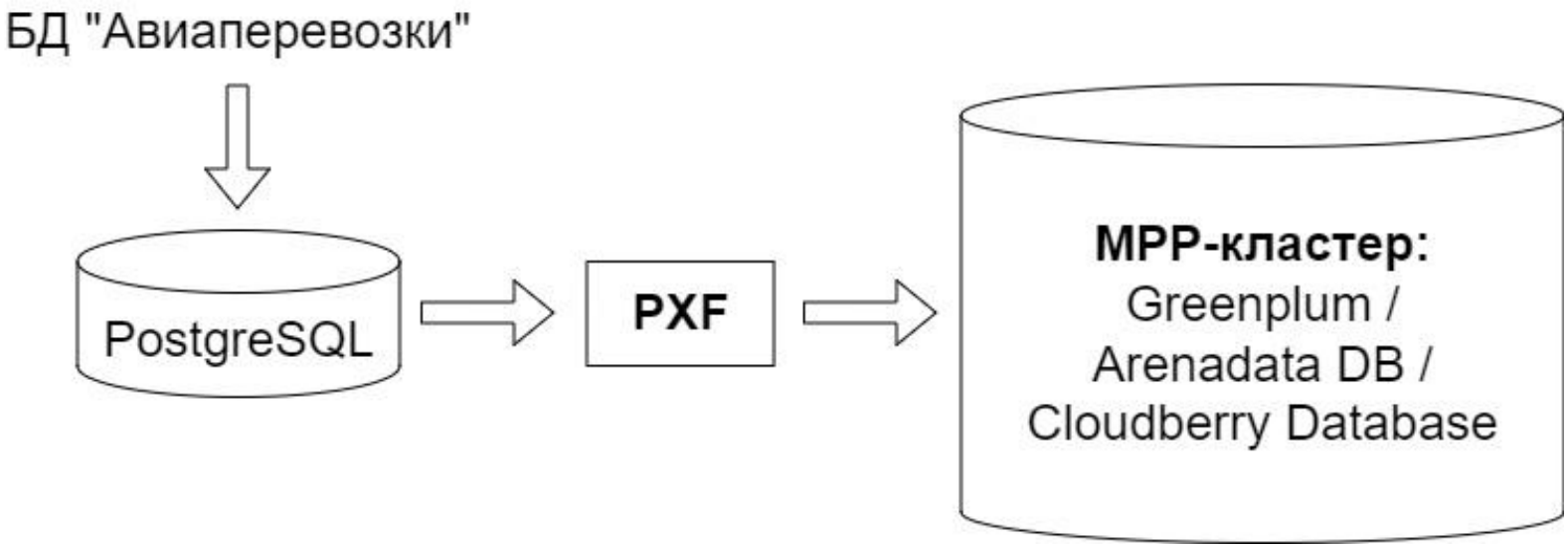


Схема кластеров



Версии PostgreSQL в MPP-кластерах

Кластер	Версия кластера	Версия PostgreSQL
Greenplum	6.23, 6.27	9.4
Arenadata DB	7.2	12.12
Cloudberry Database	1.5.1, 1.6	14.4

Варианты оптимизаторов (планировщиков): GPORCA, Postgres.

Источники по методам оптимизации

1. Лекция «оптимизация запросов» и другие лекции
2. Документация по Greenplum: https://techdocs.broadcom.com/
3. Документация по Arenadata: https://docs.arenadata.io
4. 5 лайфхаков оптимизации SQL-запросов в Greenplum
   https://habr.com/ru/companies/rostelecom/articles/442758/

Методы оптимизации запросов MPP (1)

1. Распределение данных, соединения по ключам дистрибуции. Для ключей дистрибуции, по которым будут соединения, использовать одинаковые типы данных.
2. Партиционирование, partition elimination
3. Использование append optimized таблиц, использование колоночной ориентации
4. Использование distributed replicated для маленьких таблиц (справочников)
5. Использование unlogged таблиц
6. Использование temporary таблиц для хранения промежуточных результатов вычислений

Методы оптимизации запросов MPP (2)

7. Добиваться равномерного распределения данных между сегментами. Избегать skew – перекосов.
8. По возможности, избегать переноса недостающих данных с одного сегмента на другой: broadcast motion и redistribute motion.
9. По возможности, избегать операций сортировки (order by в запросе, sort в плане запроса)
10. Поддержание статистики в актуальном состоянии, регулярный сбор
11. Использование индексов
12. Управление оптимизацией с помощью параметров

Сложные и интересные случаи оптимизации

1. Использование Nested Loop:
   - одна из соединяемых таблиц – маленькая,
   - соединение не по равенству: >, <, >=, <=, …
2. Сканирование индексов – Index Scan
3. Запросы, для оптимизации которых требуются специальные виды статистики
4. Использование параметров для управления оптимизацией
5. Сортировки в оконных функциях
6. Запросы с InitPlan
7. Комбинированная группировка

Что получилось

1. Развернуты кластеры: Greenplum, Arenadata DB, Cloudberry Database
2. Базы данных заполнены специальными данными для воспроизведения сложных случаев оптимизации запросов
3. Из разных источников собраны методы оптимизации запросов
4. Произведен разбор некоторых сложных запросов и планов их выполнения
5. Сделаны выводы

Выводы

1. MPP-кластеры отстают от БД PostgreSQL в оптимизации запросов
2. Чем больше версия MPP-кластера и БД, входящих в его состав, тем больше возможностей по оптимизации запросов
3. Все запросы, которые можно выполнить с помощью Sec Scan, MPP-кластеры выполняют с помощью стой операции и обеспечивают хорошее время выполнения за счет распараллеливания по сегментам
4. Методы оптимизации, собранные при выполнении данной проектной работы, позволят ускорять запросы в предстоящих проектах
5. Полученные навыки создания MPP-кластеров позволят эффективно развертывать стенды для разработки и тестирования

Вопросы и рекомендации

\+ если есть вопросы

— если вопросов нет

Спасибо за внимание!