



ОНЛАЙН-ОБРАЗОВАНИЕ

Онлайн-образование

Не забыть включить запись!





Меня хорошо видно && слышно?

Ставьте ☐+, если все хорошо
Напишите в чат, если есть проблемы





Работа с большими объёмами данных

Курочкин Константин
«Medindex»

Правила вебинара



Активно участвуем



Задаем вопрос в чат



Off-topic обсуждаем в telegram



Вопросы вижу в чате, могу ответить не сразу

Маршрут вебинара

Обсудим термин «большие данные»



Изучим особенности работы с ними



Попробуем их загрузить в PostgreSQL



Поэкспериментируем

Цели вебинара | После занятия вы сможете

1

Понимать природу больших данных и знать места их обитания

2

Загружать большие данные в PostgreSQL

3

Иметь представление о том, как работать с большими данными в PostgreSQL

The image features a high-angle, aerial view of a dense urban landscape, likely New York City, with numerous skyscrapers and buildings. The entire image is overlaid with a semi-transparent blue and green gradient. A network of thin, light blue lines connects various points across the gradient, creating a digital or data-like aesthetic. The text "Что такое большие данные?" is centered in the middle of the image in a white, sans-serif font.

Что такое большие данные?

Что такое большие данные?

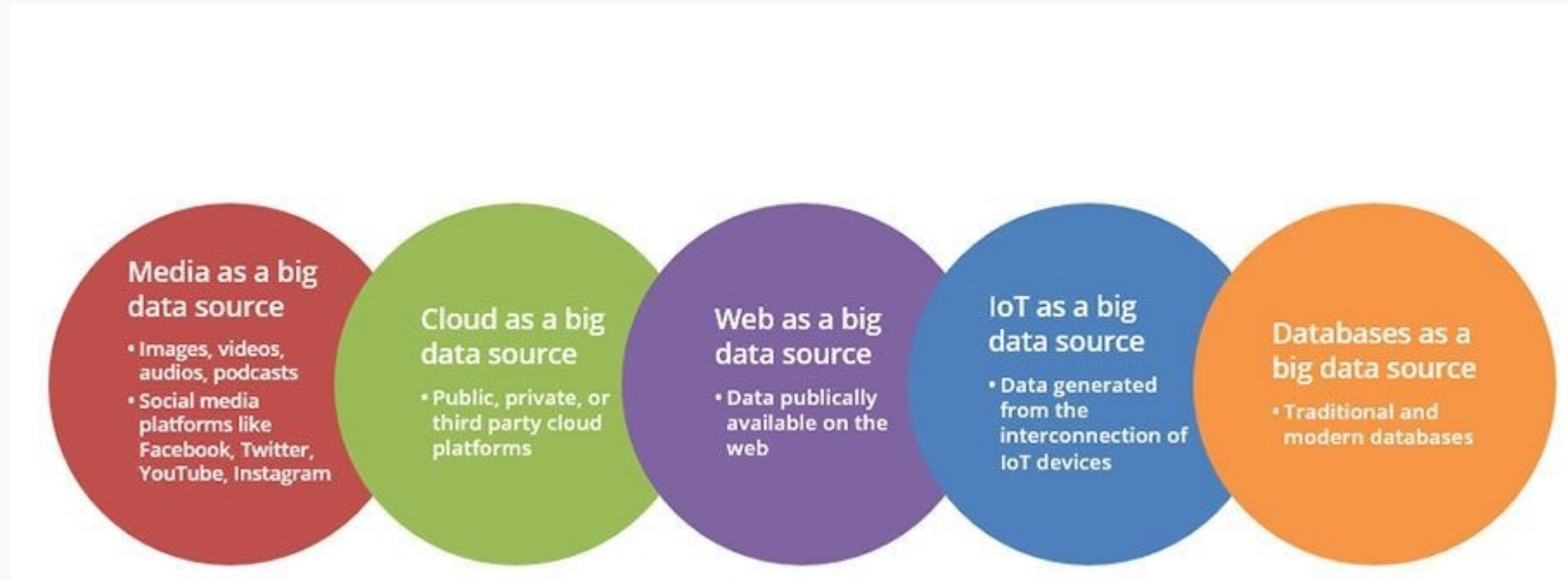
Большие данные ([англ.](#) big data)

- серия подходов, инструментов и методов обработки структурированных и [неструктурированных данных](#) огромных объёмов и значительного многообразия
- для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста, распределения по многочисленным узлам [вычислительной сети](#)
- сформировавшихся в конце [2000-х годов](#), альтернативных традиционным [системам управления базами данных](#) и решениям класса Business Intelligence

Примеры больших данных

- логи поведения пользователей в интернете
- GPS-сигналы от автомобилей для транспортной компании
- данные, снимаемые с датчиков в большом адронном коллайдере
- информация о транзакциях всех клиентов банка
- информация о всех покупках в крупной ритейл сети и т.д.

Откуда берутся большие данные?



Принципы работы с большими данными

горизонтальная масштабируемость

- сотни и даже тысячи узлов

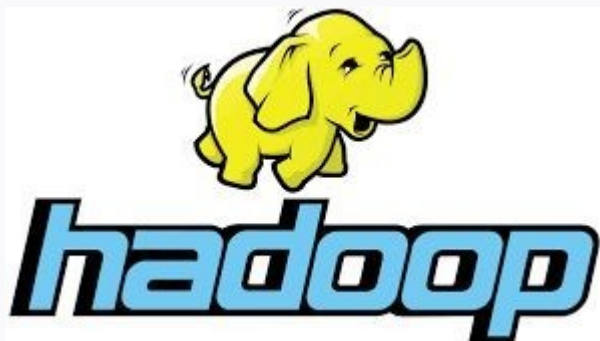
отказоустойчивость

- постоянное добавление и удаление узлов
- десятками и даже сотнями

локальность данных

- данные распределены по множеству узлов
- шардирование
- дублирование

Примеры систем для работы с большими данными



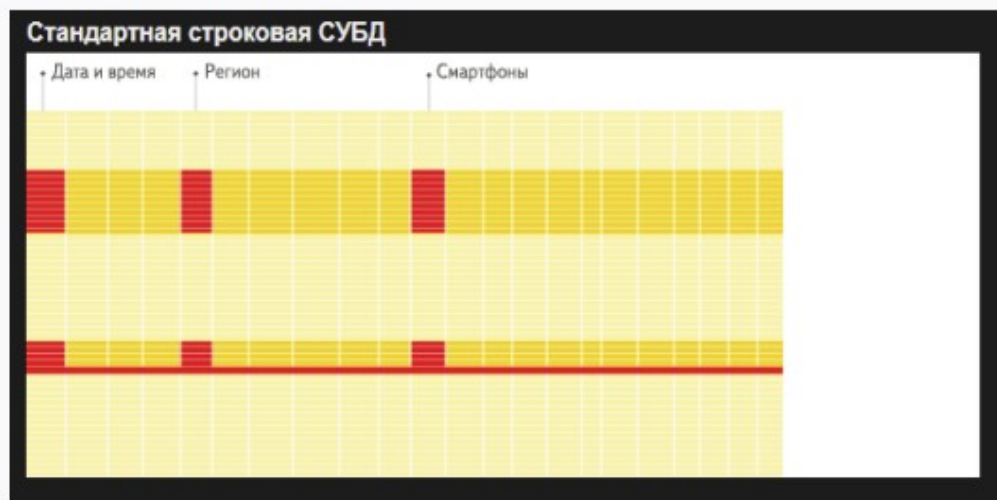
Google
BigQuery



ClickHouse

VERTICA

Примеры систем для работы с большими данными



А где здесь PostgreSQL?

его нет

- PostgreSQL не предназначен для работы с большими данными*
- как впрочем любая другая транзакционная СУБД
- а почему нет (давайте подумаем)?

- * однако он является прекрасным источником для них



Почему нет?

транзакции и WAL

- механизм транзакций мешает работе с большими данными
- механизм WAL мешает работе с большими данными

горизонтальное масштабирование

- отсутствие эффективного механизма горизонтального масштабирования
- отсутствие эффективного механизма шардирования данных

Тогда зачем?

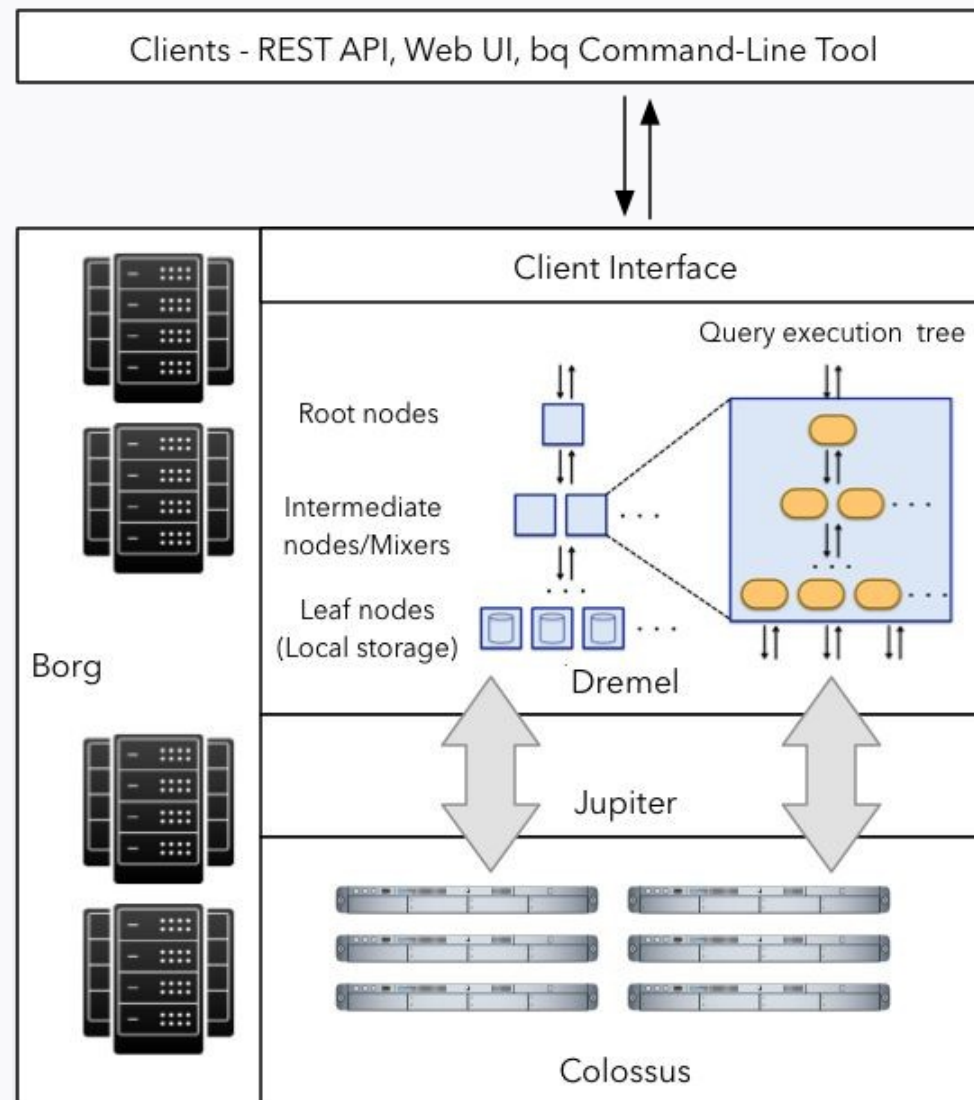
- Если все таки придется столкнуться
- Если объем вашей OLTP БД вырастет до объема BigData
- Ну и надо же на чем то тренироваться ;)

Google BigQuery

- Закрытый продукт Google
- Основан на распределенной файловой системе Colossus
- Колоночно ориентированный
- Нет индексов и транзакций
- Data warehouse
- Полностью SaaS
- SQL интерфейс



Google BigQuery



Практика

- демо датасет `bigquery-public-data:chicago_taxi_trips.taxi_trips`
- что интересного внутри
- сколько записей и какой размер
- выгружаем CSV в GCS



Загрузка больших данных в PostgreSQL



Хорошие способы загрузки данных в PostgreSQL

Foreign Data Wrappers ([FDW](#))

- Foreign Data Wrappers in PostgreSQL and a closer look at [postgres fdw](#)
- SQL [copy](#)
- [pgloader](#)
- [pg_bulkload](#)

Перед загрузкой стоит:

- Отключить автокомит
- Убрать индексы
- Убрать внешние ключи
- Увеличить параметр `maintenance_work_mem` (скорее поможет после загрузки)
- Увеличить `max_wal_size`

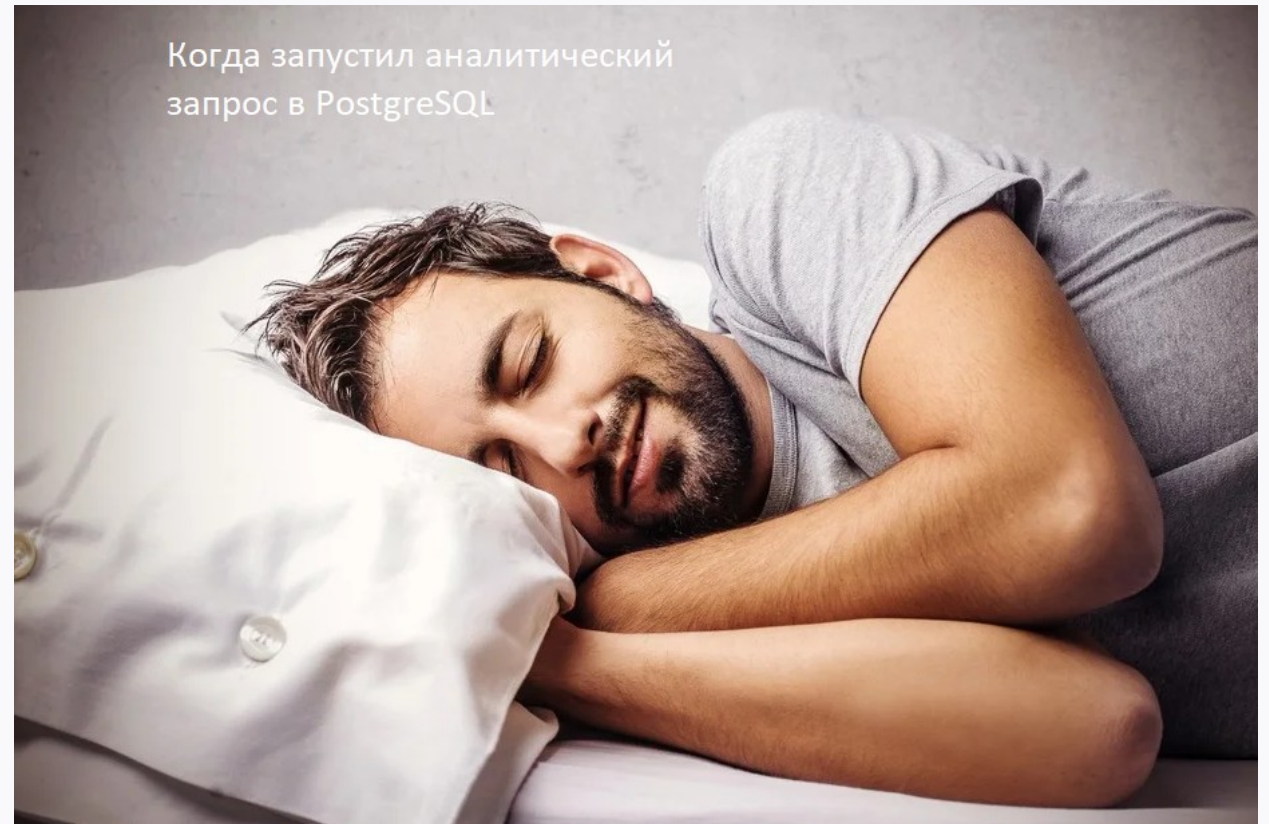
Более подробно [здесь](#)

Рубрика «Очумелые ручки»

Возьмём параметры из `pg_tune` для OLAP и посмотрим, справится ли наш сервер с запросом

А что делать?

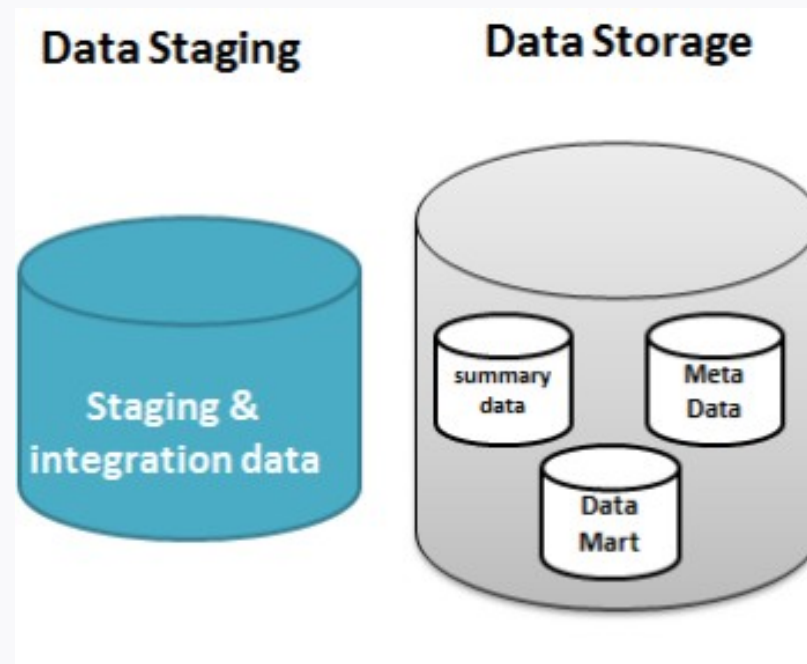
- Партиционирование
- Витрины
- Модель данных
- Правильное индексирование



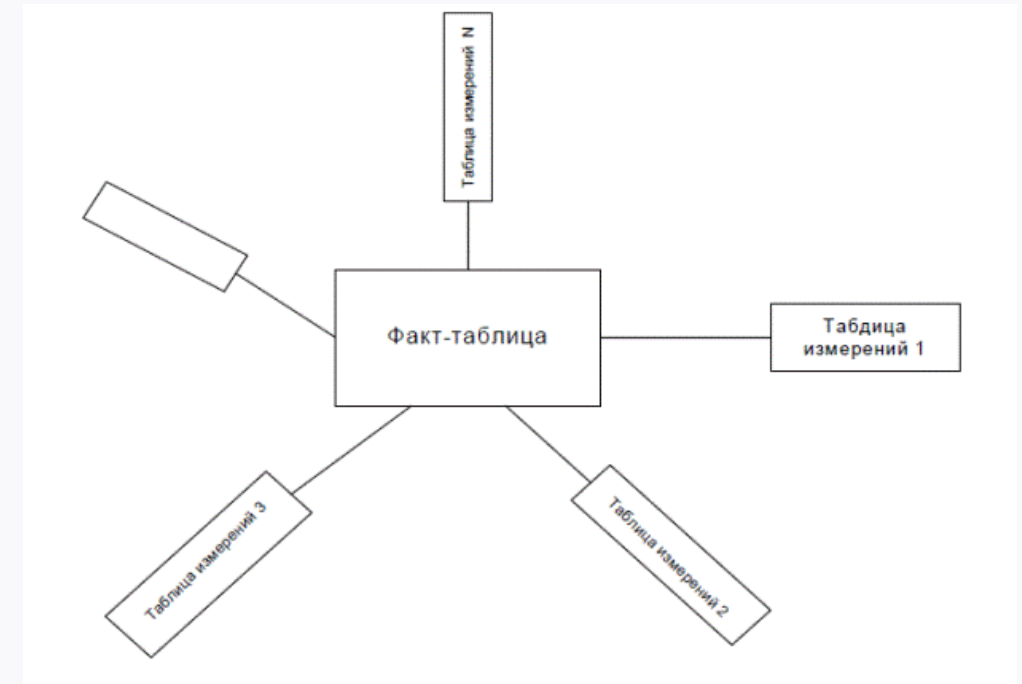
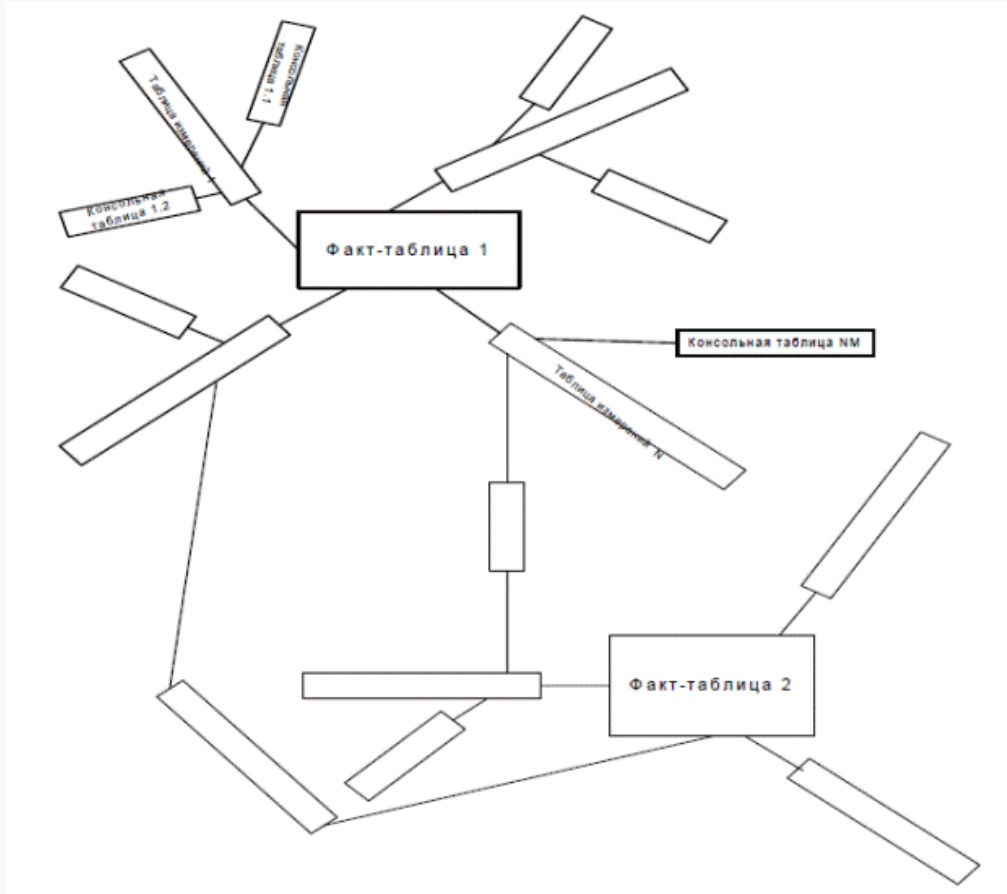
Партиционирование

```
CREATE TABLE SOME_TABLE (  
  id int generated by default as identity,  
  /*some_columns*/  
  some_date date)  
partition by range (some_date);
```


Витрины

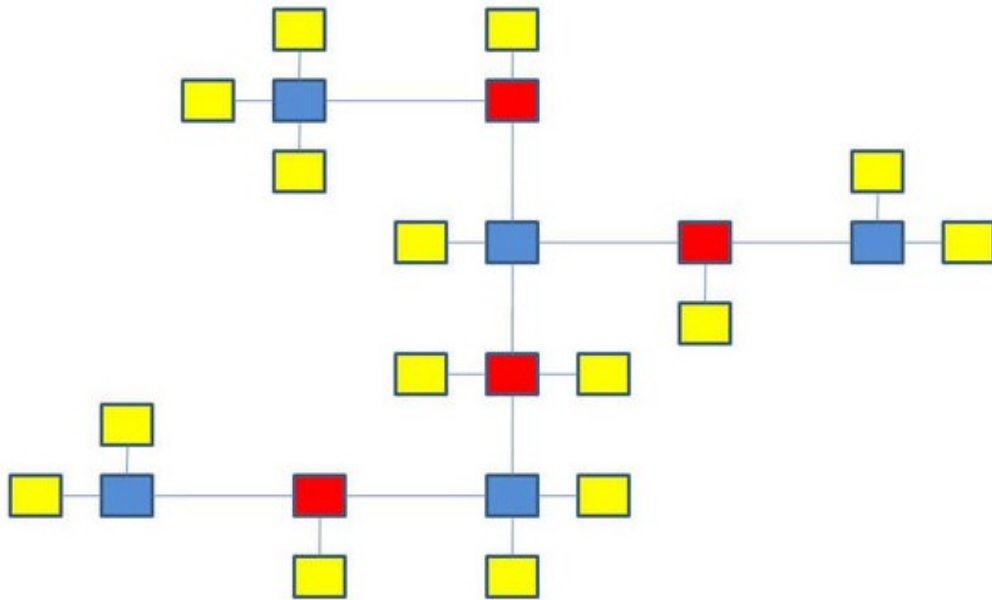


Модель данных



Модель данных DataVault

Hubs / Links / Satellites



Хаб (Hub)

Единичная бизнес-сущность:
Клиент, Договор, Поставка и т.п.

- Уникальный бизнес-ключ (набор ключей)
- Суррогатный ключ (рекомендуется использовать хеш бизнес-ключей)
- Дата-время загрузки
- Источник записи

Связь (Link)

Связь между сущностями (хабами).

- Суррогатные ключи связываемых сущностей (их может быть более 2!)
- Дата-время загрузки
- Источник

Спутник (Satellite)

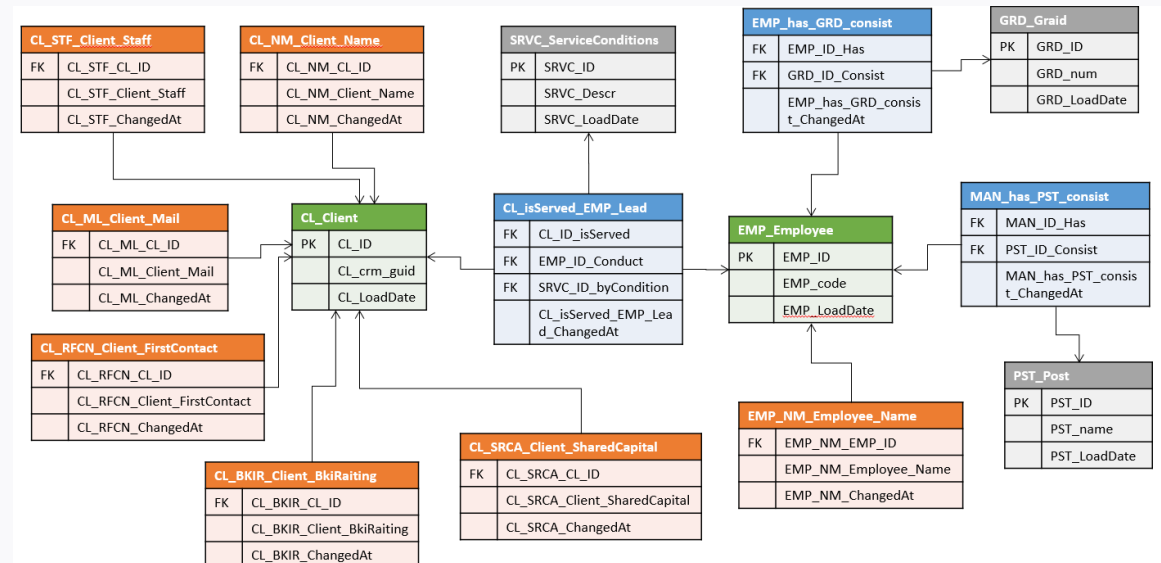
Описательные атрибуты сущностей (хабов) и связей.

- Суррогатный ключ родителя
- Значения атрибутов
- Дата-время начала действия версии
- Дата-время загрузки
- Источник

Модель данных Anchor modelling

- 6 НФ
- Ещё больше джойнов ☺
- <http://www.anchormodeling.com/postgresql-12-and-editing-en-masse/>

Важное отличие от Data Vault: Хаб – список бизнес-ключей,
Якорь – список суррогатов!

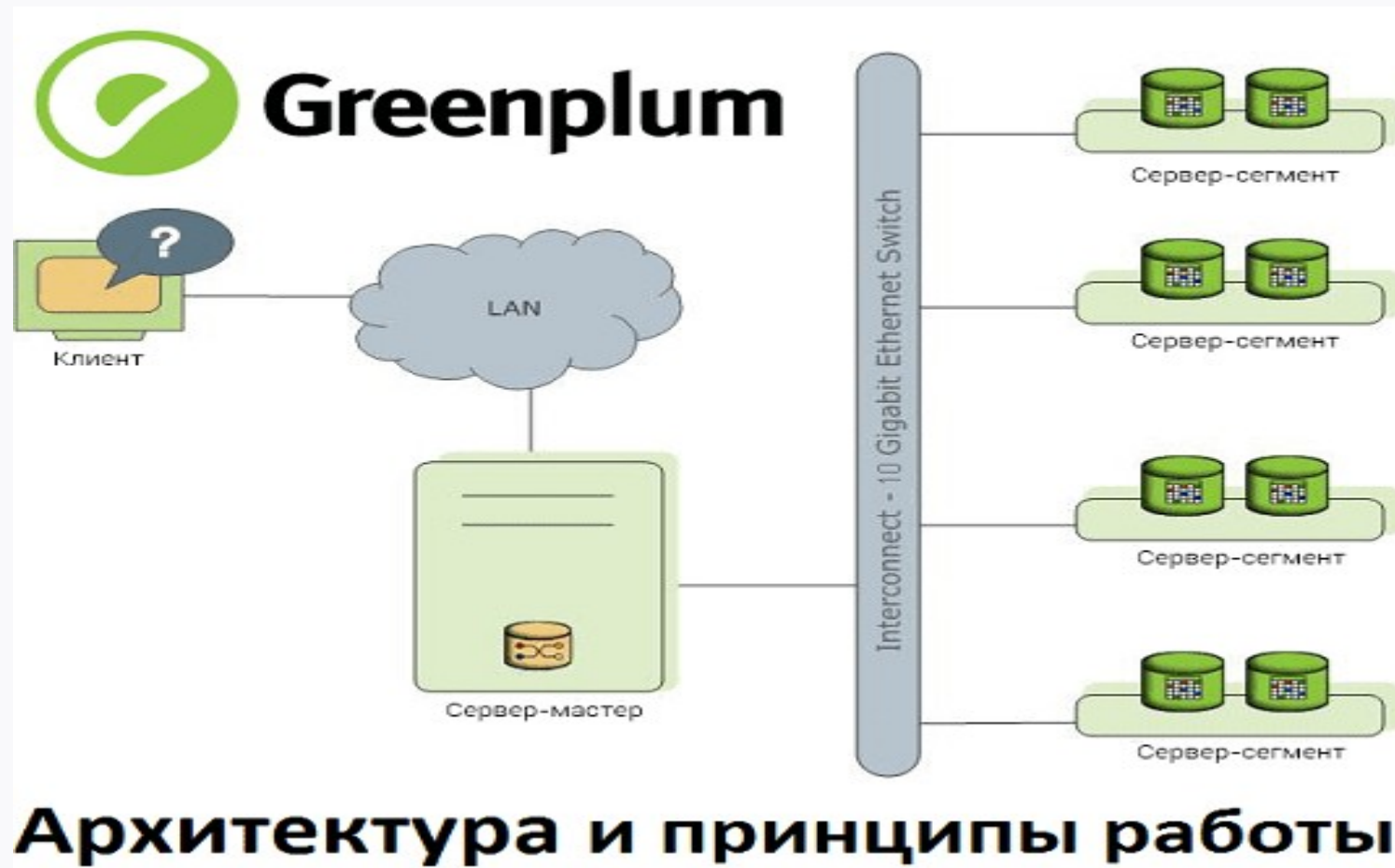


А можно мне вот то, зелёненькое?

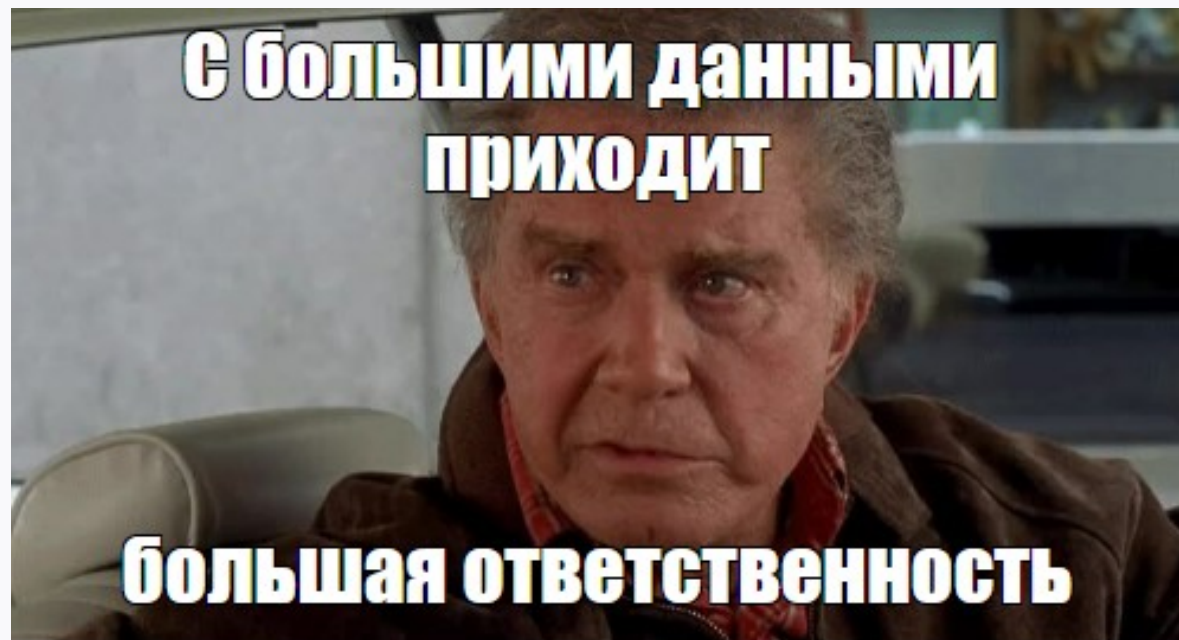
- Бесплатная
- Колоночная
- MPP
- На основе PostgreSQL 9.3
- <https://greenplum.org/>



А можно мне вот то, зелёное?



Послесловие




Рефлексия




Отметьте самый не раскрытый, по вашему мнению пункт



Будете ли вы работать с большими данными в PostgreSQL? Почему? 😊

The background of the image is an aerial photograph of a dense city skyline, likely New York City, with numerous skyscrapers. The image is overlaid with a semi-transparent blue layer. On this blue layer, there is a white network pattern consisting of dots connected by thin lines, resembling a molecular or digital structure. The text is centered on this blue layer.

Заполните, пожалуйста,
опрос о занятии по ссылке в чате
<https://otus.ru/polls/63773/>



До новых встреч!
Приходите на следующие занятия

Курочкин Константин
Medindex