Тинькофф

SQL
Применение знаний

Tinkoff.ru

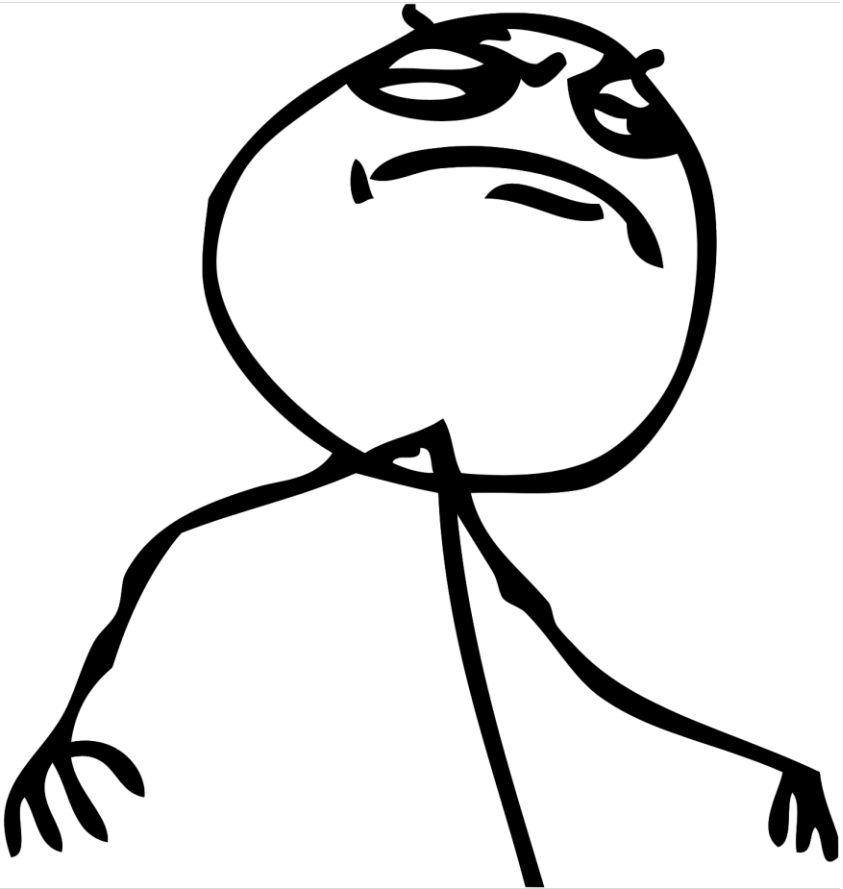
2020

Предпосылки

Кто владеет информацией, тот владеет миром.

Н.М. Ротшильд

Предпосылки



Это Валера.

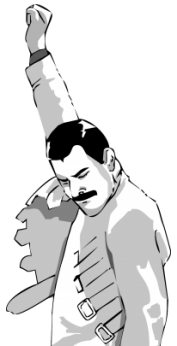
Он собрал много информации обо всех клиентах банка: кто и когда совершал какую финансовую операцию, на какую кнопку в интернет-банке кликал, куда перемещал курсором мыши во время посещения сайта, кому переводил деньги. Теперь у него 100ТВ информации, и он считает себя победителем.

Предпосылки

Это Маша.

В отличие от Валеры, она **структурировала** данные. Все данные обработаны, систематизированы и доступны для удобной аналитики. У неё всего 1ТВ данных, но на них построена высокодоступная и производительная модель данных. Бизнес принимает эффективные операционные решения и корректирует свою стратегию на рынке благодаря Маше.

Маша – настоящий победитель.



Предпосылки

> Кто владеет информацией, тот владеет миром.
>
> Н.М. Ротшильд

- Владеть инфомацией – мало. Необходимо уметь ее анализировать.
  - Анализировать – превратить в знания и последующие решения.
  - Обладая меньшими данными, можно извлекать из них бОльшие знания и принимать более верные решения.
- Растут не только объемы информации, представляющей аналитическую ценность, но и скорость их роста.
  - Мы хотим анализировать клиента по его «следам» в интернете.
  - Каждый сервис дает ценную информацию о поведении клиента.
- Требуемая скорость принятия решений растет.
  - Падение тренда конверсии посетителя сайта в заявку – теряем клиентов!

Предпосылки

Чтобы анализировать данные, нужно уметь:

- хранить данные:
  - структурированно (должен быть порядок);
  - надежно (данные не должны теряться);
  - безопасно (разграничение прав доступа);
- объединять данные из разных массивов;
- создавать производные от данных;
- иметь удобный доступ к данным:
  - программный;
  - ручной.

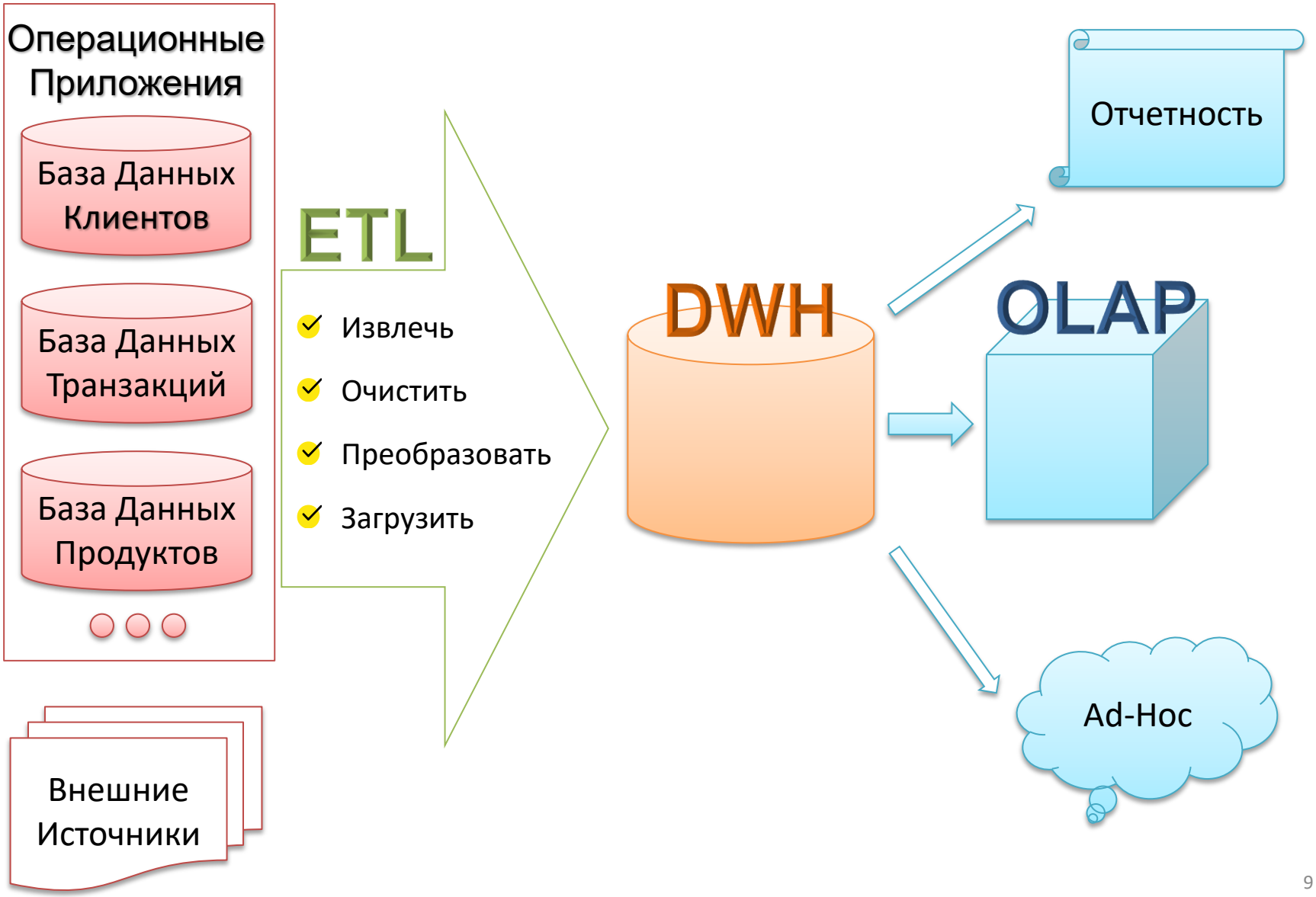
Решение

Аналитическая платформа

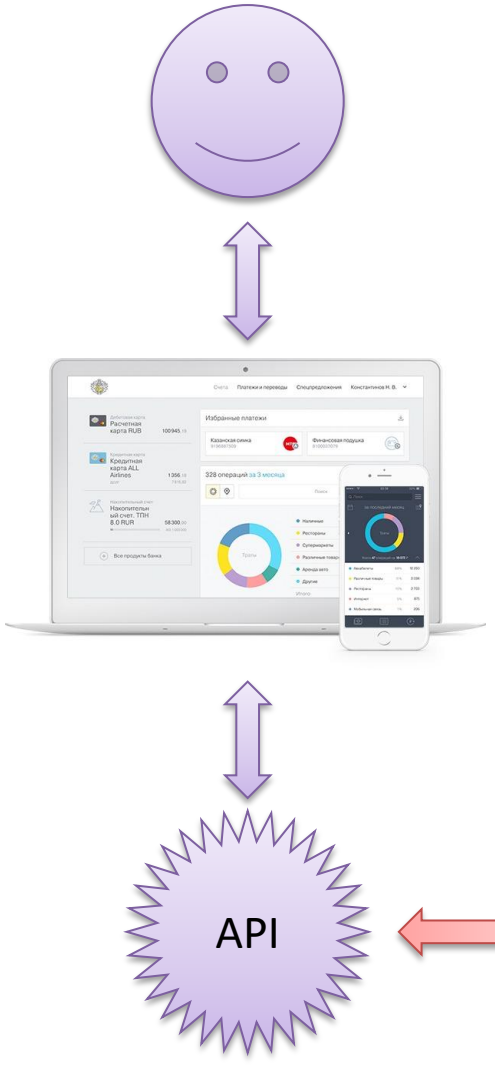
- Реализовать функциональность:
  - Ad hoc аналитика;
  - анализ данных в режиме реального времени;
  - регламентная отчетность.
- Дать возможность:
  - быстрый и удобный доступ к данным;
  - гибкая визуализация процесса и результатов анализа.

Business Intelligence – набор технологий, методологий, инструментов и мероприятий, направленных на перевод необработанной информации (сырых данных) в осмысленную, удобную форму (бизнес-знания).

Архитектура Business Intelligence



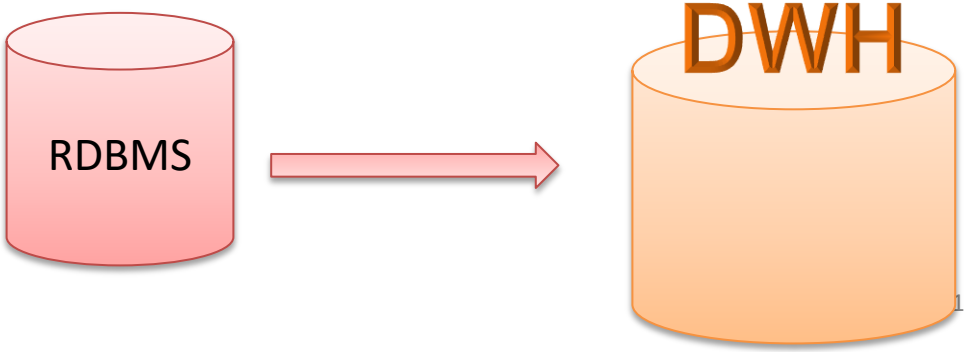
Приложения OLTP



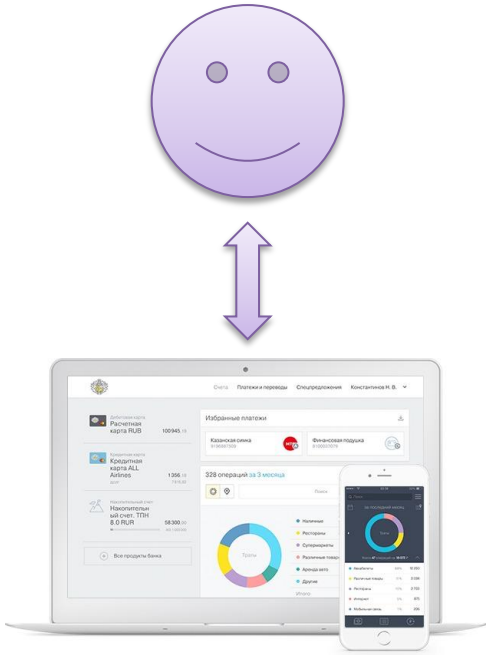
- Много одновременных операций.
- Операции затрагивают один или малое количество объектов СУБД.
- Запросы выполняются очень быстро.
- В одной операции не бывает массового изменения данных.
- Создаются новые данные.
- Данные могут удаляться.
- Соблюдается целостность.
- Хранятся технические атрибуты.

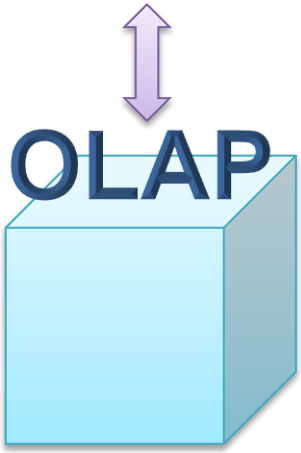
Приложения OLAP

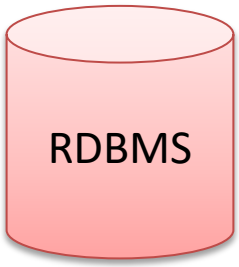
- Небольшое количество одновременных операций.
- Операции затрагивают одновременно большое количество объектов СУБД.
- Запросы могут выполняться сколь угодно долго.
- Изменения данных обычно массовые.
- Новые данные не создаются.
- Данные не могут удаляться.
- Создаются новые расчетные атрибуты.
- Создается историчность значений атрибутов.
- Данные более статичны.

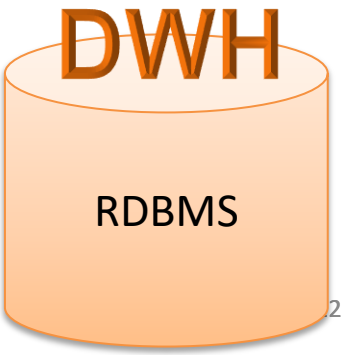


OLTP и OLAP - это RDBMS









Принципы организации Хранилищ Данных

- Проблемно-предметная ориентация
- Интегрированность
- Информация в хранилище не создается, а только накапливается извне
- Историчность данных

Модель данных

Модель данных – модель «сущность-связь» (ER-модель), описывающая на нескольких уровнях набор взаимосвязанных сущностей, которые отражают потребности бизнеса в аналитике и отчетности.

- Концептуальная модель данных
- Логическая модель данных
- Физическая модель данных

Коцептуальная модель

Концептуальная модель данных – описание основных сущностей и отношений между ними.

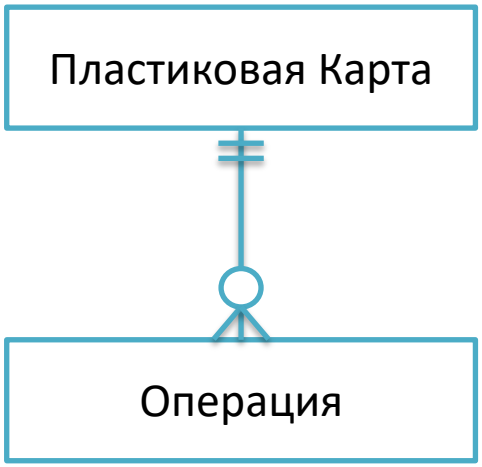
Как проектируется концептуальная модель?

- Исследование бизнеса
- Выделение ключевых объектов, которыми оперирует бизнес
- Определение логических отношений между объектами

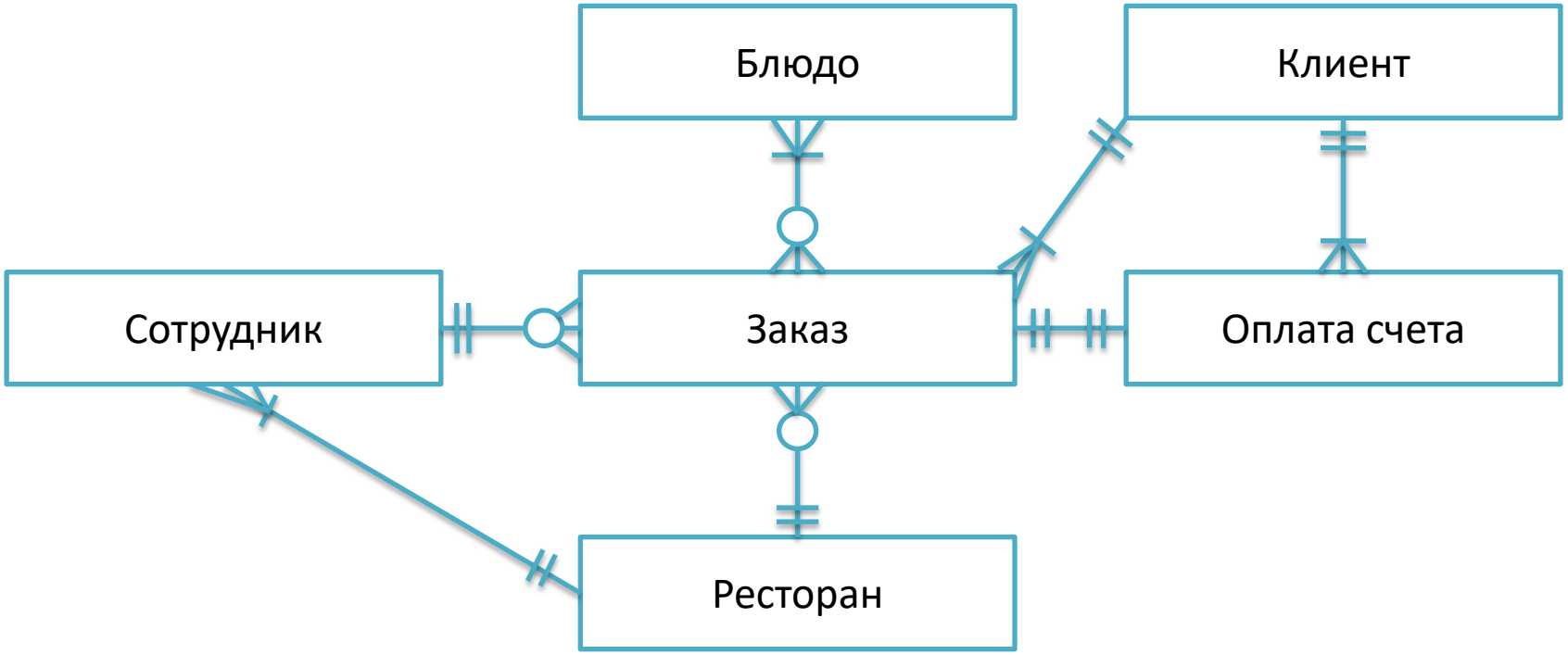
Коцептуальная модель

Модель типа «сущность-связь». Что определяем?

- Сущности
- Связи между ними



Концептуальная модель ХД сети ресторанов



Омонимы

- Омоним- слово, однозвучное с другим, но отличное от него по значению.

Омонимы в деловой среде

Активный студент - ?

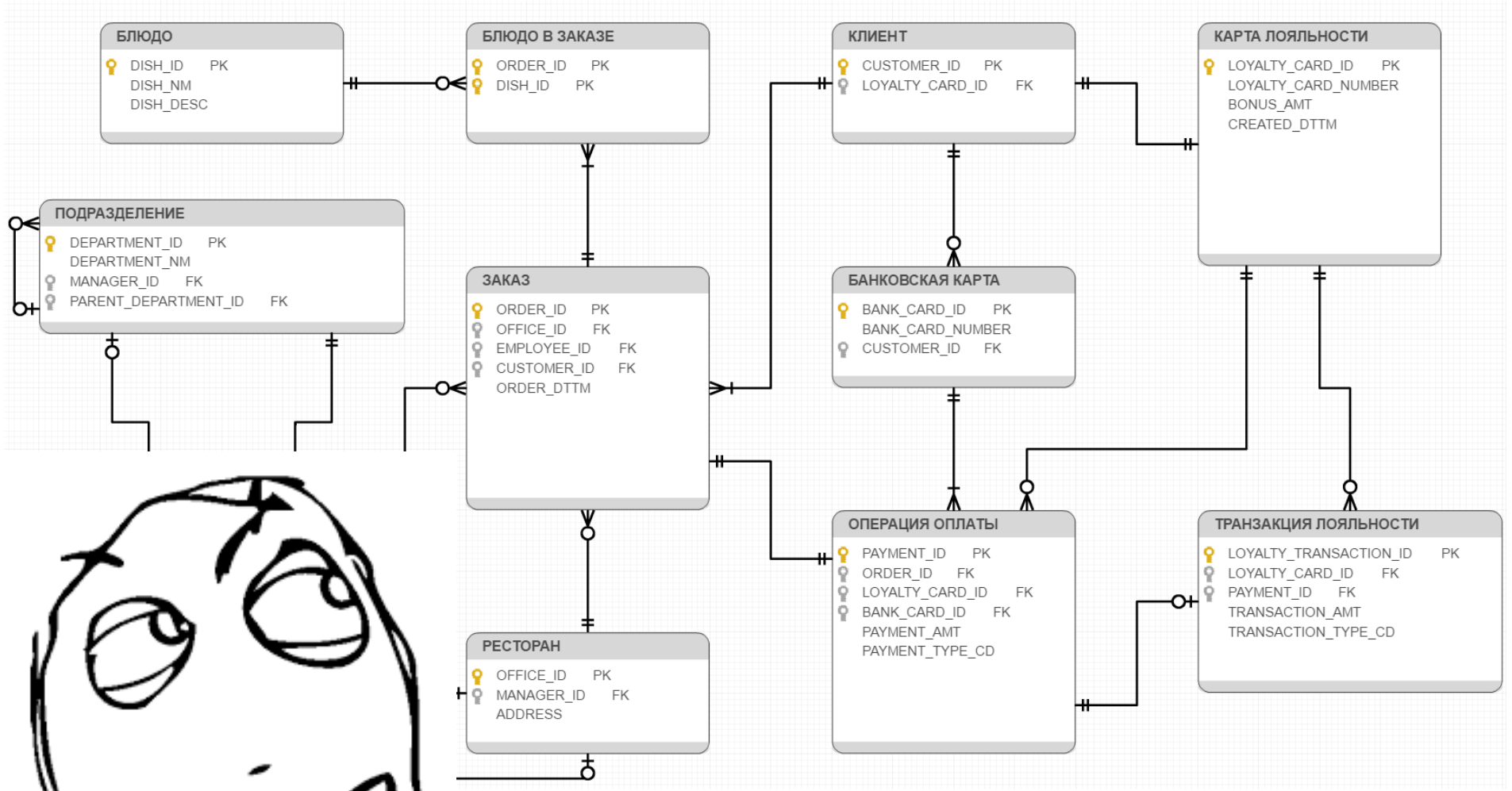
- Ходит на пары?
- Сдал сессию на 4 и 5?
- Участвует в КВН?
- Волонтёрит иногда?
- Ходит ли на выборы?
- …
- Ведет активный образ жизни?...

Логическая модель

Логическая модель данных – расширение концептуальной модели данных путем определения для сущностей их атрибутов, описаний и ограничений, частично уточняет состав сущностей и взаимосвязи между ними.

- Разрешает выход за рамки концептуальной модели
- Является прототипом будущей физической модели
- Не учитывает специфику какой-либо конкретной СУБД

Логическая модель ХД сети ресторанов



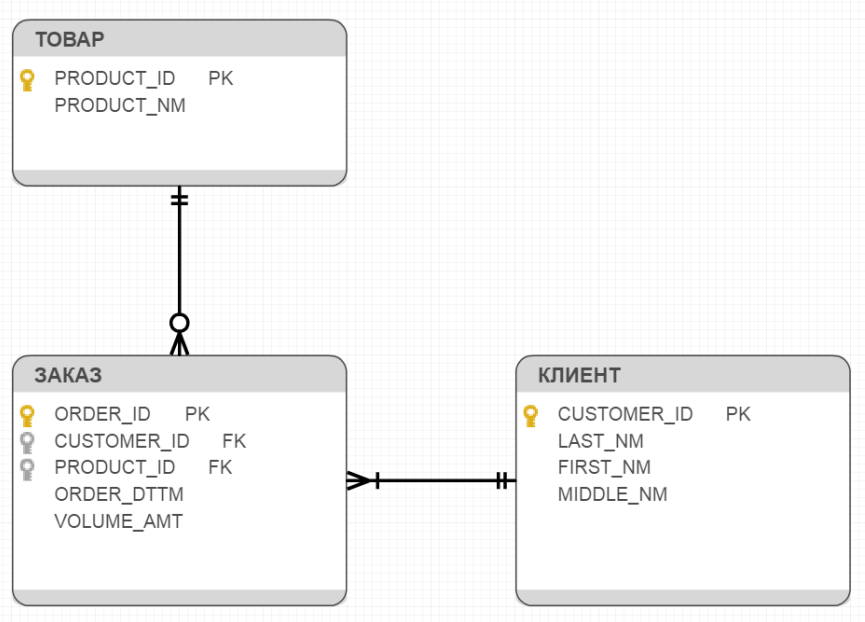
Важно знать

Логическая модель данных – ключевая ценность, которую несет для бизнеса Хранилище Данных.

Реляционные базы данных

product_id	product_nm
58546918	НАЗВАНИЕ
5722434	НАЗВАНИЕ
78321383	НАЗВАНИЕ
78155977	НАЗВАНИЕ
78571781	НАЗВАНИЕ
82231929	НАЗВАНИЕ
3393773	НАЗВАНИЕ
78901430	НАЗВАНИЕ
79587792	НАЗВАНИЕ

order_id	customer_id	product_id	order_dttm	volume_amt
21454279	130849138	78321383	2016-03-18 00:0...	5
23417066	33728505	79587792	2016-04-15 00:0...	52
21141517	556757	78901430	2016-03-30 00:0...	219
1588009	2903087	3393773	2011-06-17 00:0...	204
21492701	131091302	78571781	2016-03-23 00:0...	16
2871788	5117420	5722434	2012-06-01 00:0...	139
21193031	20751639	82231929	2016-05-30 00:0...	234
11386045	112898869	58546918	2014-07-05 00:0...	97
23114353	115711210	78155977	2016-03-19 00:0...	335



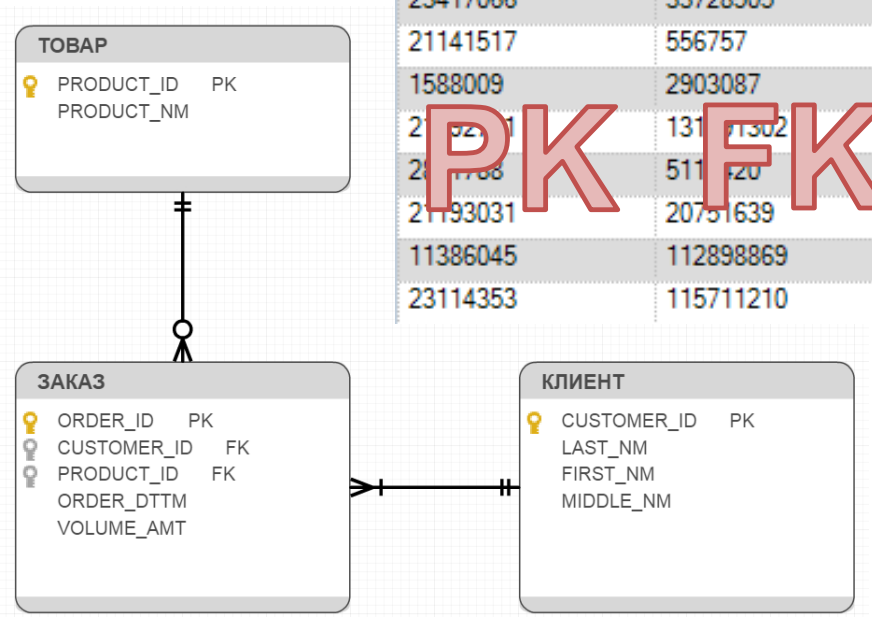
customer_id	last_nm	first_nm	middle_nm
112898869	ФАМИЛИЯ	ИМЯ	ОТЧЕСТВО
115711210	ФАМИЛИЯ	ИМЯ	ОТЧЕСТВО
130849138	ФАМИЛИЯ	ИМЯ	ОТЧЕСТВО
5117420	ФАМИЛИЯ	ИМЯ	ОТЧЕСТВО
33728505	ФАМИЛИЯ	ИМЯ	ОТЧЕСТВО
20751639	ФАМИЛИЯ	ИМЯ	ОТЧЕСТВО
131091302	ФАМИЛИЯ	ИМЯ	ОТЧЕСТВО
2903087	ФАМИЛИЯ	ИМЯ	ОТЧЕСТВО
556757	ФАМИЛИЯ	ИМЯ	ОТЧЕСТВО

Первичные и внешние ключи

- Первичный ключ – поле или набор полей, идентифицирующих строку в таблице:
  - недопустимо отсутствие значения;
  - все значения уникальны.
- Внешний ключ – поле или набор полей, устанавливающих связь между данными в двух таблицах. В общем случае не имеет логических ограничений, как первичный ключ.

order_id	customer_id	product_id	order_dttm	volume_amt
21454279	130849138	78321383	2016-03-18 00:0...	5
23417066	33728505	79587792	2016-04-15 00:0...	52
21141517	556757	78901430	2016-03-30 00:0...	219
1588009	2903087	3393773	2011-06-17 00:0...	204
2[illegible]	131[illegible]	78[illegible]	2016-03-23 00:0...	16
2[illegible]	511[illegible]	57[illegible]	2012-06-01 00:0...	139
21193031	20751639	82231929	2016-05-30 00:0...	234
11386045	112898869	58546918	2014-07-05 00:0...	97
23114353	115711210	78155977	2016-03-19 00:0...	335

PK FK FK

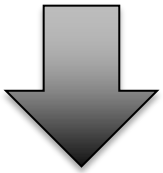


Определение 1NF

Таблица находится в **первой нормальной форме** тогда и только тогда, когда любая ячейка (любое поле любой строки) содержит только одно логическое значение.

Свойство атомарности.

Клиент				
Ключ	**Фамилия**	**Имя**	**Отчество**	**Детали**
1	ИВАНОВ	ИВАН	ИВАНОВИЧ	Родился в 1985 году, получил высшее образование в МГУ

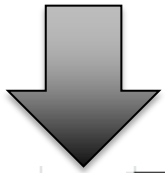


Клиент						
Ключ	**Фамилия**	**Имя**	**Отчество**	**Год рождения**	**Образование**	**ВУЗ**
1	ИВАНОВ	ИВАН	ИВАНОВИЧ	1985	Высшее	МГУ

Определение 2NF

Таблица находится во **второй нормальной форме** тогда и только тогда, когда она находится в первой нормальной форме и значение в каждом поле, не принадлежащем первичному ключу, зависит от полного набора полей в первичном ключе.

Музыкальные диски				
Название группы	**Название СД-диска**	**Название песни**	**Автор слов**	**Композитор**
Scorpions	World Wide Live	Countdown	Klaus Meine	Matthias Jabs
Scorpions	World Wide Live	Coming Home	Rudolf Schenker	Klaus Meine
Scorpions	World Wide Live	Blackout	Rudolf Schenker	Klaus Meine
Scorpions	Blackout	Blackout	Rudolf Schenker	Klaus Meine
The Big City	Blackout	Blackout	Rudolf Schenker	Klaus Meine



Музыкальные диски		
Название группы	**Название СД-диска**	**Название песни**
Scorpions	World Wide Live	Countdown
Scorpions	World Wide Live	Coming Home
Scorpions	World Wide Live	Blackout
Scorpions	Blackout	Blackout
The Big City	Blackout	Blackout

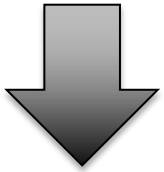
Песни			
Название группы	**Название песни**	**Автор слов**	**Композитор**
Scorpions	Countdown	Klaus Meine	Matthias Jabs
Scorpions	Coming Home	Rudolf Schenker	Klaus Meine
Scorpions	Blackout	Rudolf Schenker	Klaus Meine
The Big City	Blackout	Rudolf Schenker	Nige Roberts

Определение 3NF

Таблица находитя в **третьей нормальной форме** тогда и только тогда, когда она находится во второй нормальной форме и отсутствуют транзитивные зависимости неключевых полей от ключевых.

Иными словами, каждое неключевое поле должно содержать информацию о ключе, полном ключе и ни о чём, кроме ключа.

Счета					
Ключ	Ключ клиента	Фамилия	Имя	Отчество	Баланс
1	1	ИВАНОВ	ИВАН	ИВАНОВИЧ	250
2	1	ИВАНОВ	ИВАН	ИВАНОВИЧ	1500



Счета		
Ключ	Ключ клиента	Баланс
1	1	250
2	1	1500

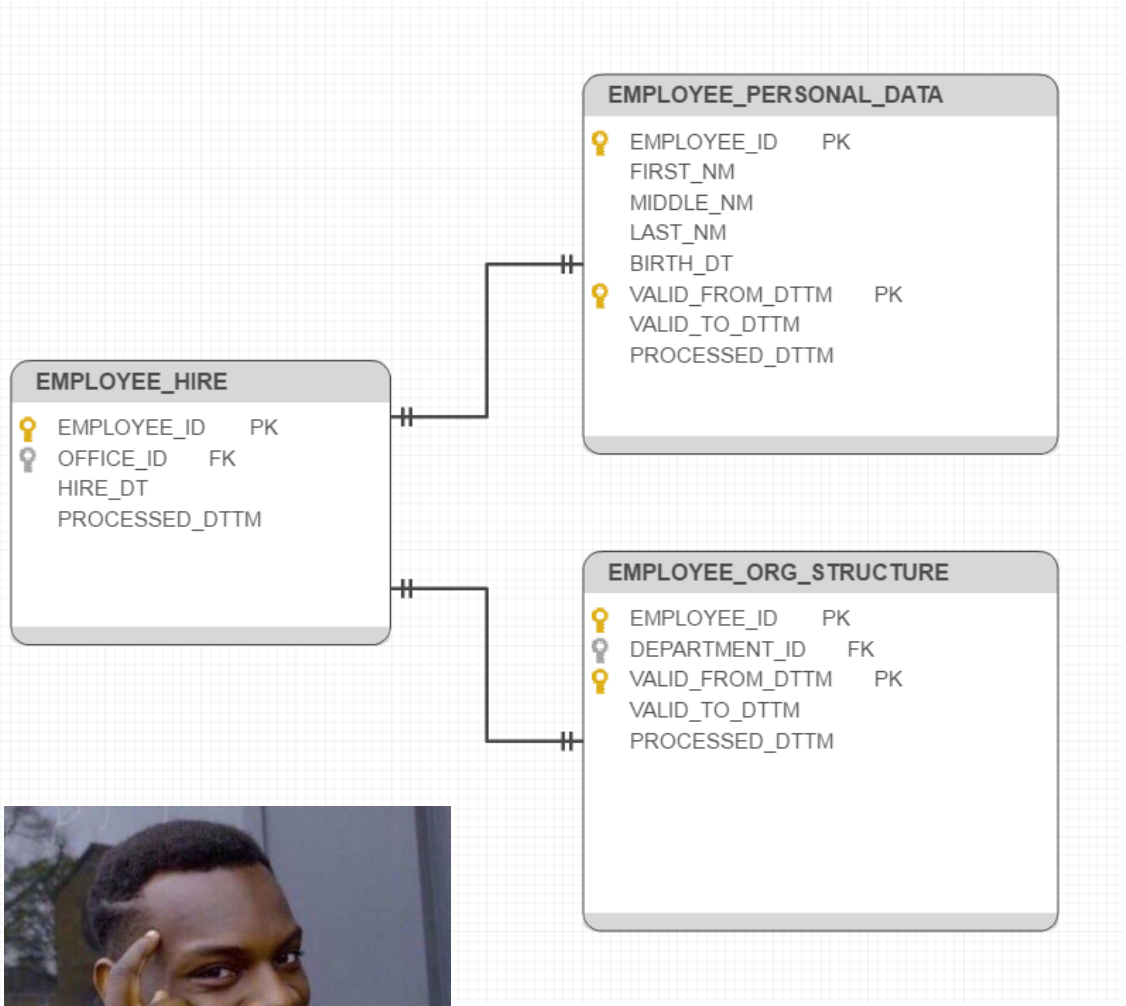
Клиент			
Ключ	Фамилия	Имя	Отчество
1	ИВАНОВ	ИВАН	ИВАНОВИЧ

Физическая модель

Физическая модель данных – описание реализации объектов логической модели на уровне конкретной базы данных с учетом всех ее особенностей.

```
CREATE TABLE DDS.IB_ACTION
(
    ib_action_rk BIGINT,
    ib_action_dttm timestamp without time zone,
    ib_action_type_cd CHARACTER VARYING(3),
    customer_rk INTEGER,
    ib_session_rk BIGINT,
    processed_dttm timestamp without time zone,
    user_rk INTEGER
)
DISTRIBUTED BY (ib_action_rk)
PARTITION BY RANGE (ib_action_dttm)
(
    PARTITION p_cmpr
    START ('1900-01-01 00:00:00'::timestamp without time zone) END ('2016-01-01 00:00:00'::timestamp without time zone)
    WITH (appendonly=true, orientation=row, compresstype=zlib, compresslevel=5),
    PARTITION p_no_cmpr
    START ('2016-01-01 00:00:00'::timestamp without time zone) END ()
    WITH (appendonly=true, orientation=row, compresstype=quicklz, compresslevel=1));
```

Физическая модель



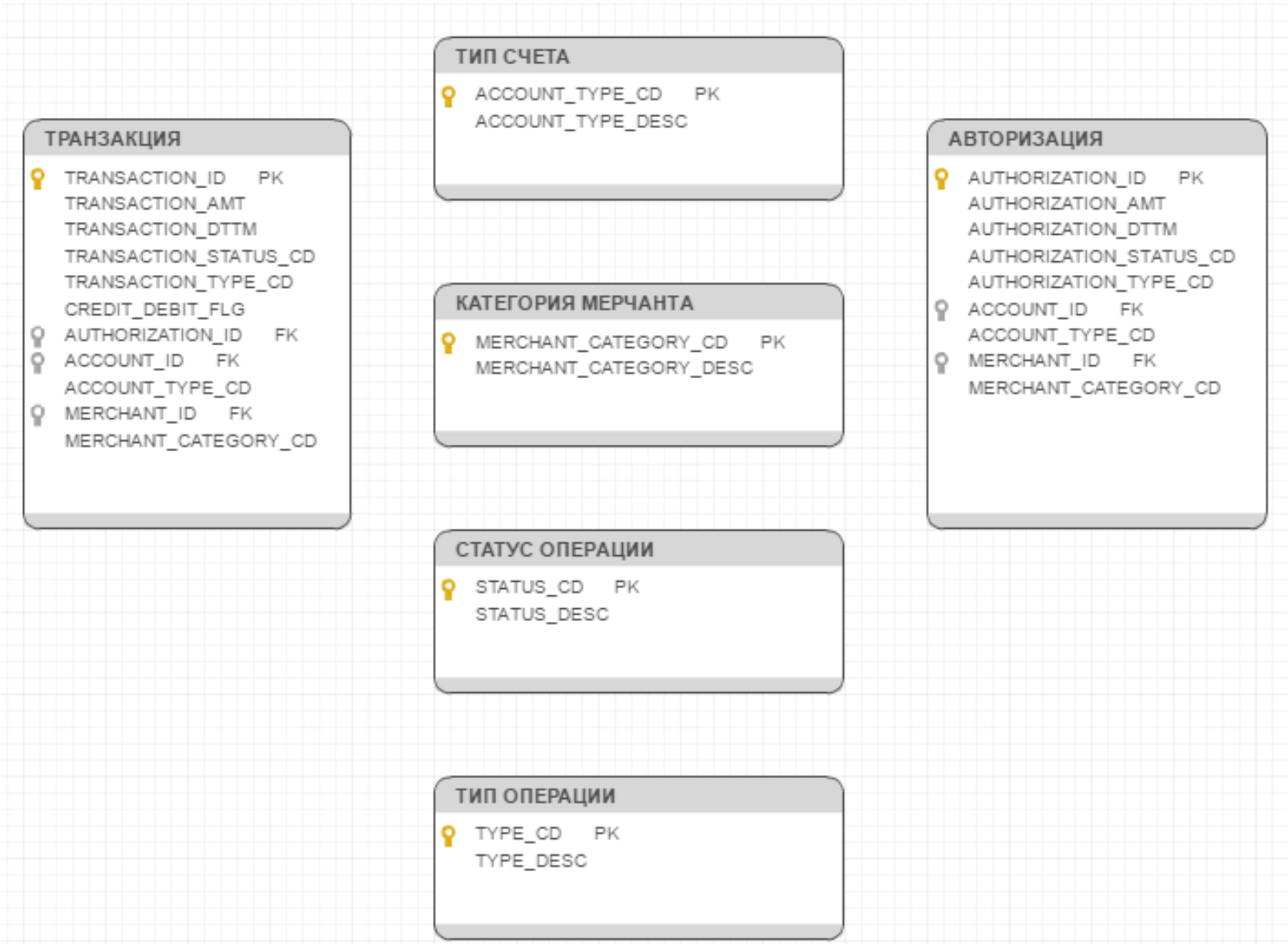
Примеры реализации физической модели данных

Наиболее существенное влияние на архитектуру физической модели данных оказывает степень нормализации данных и унификации их структуры.

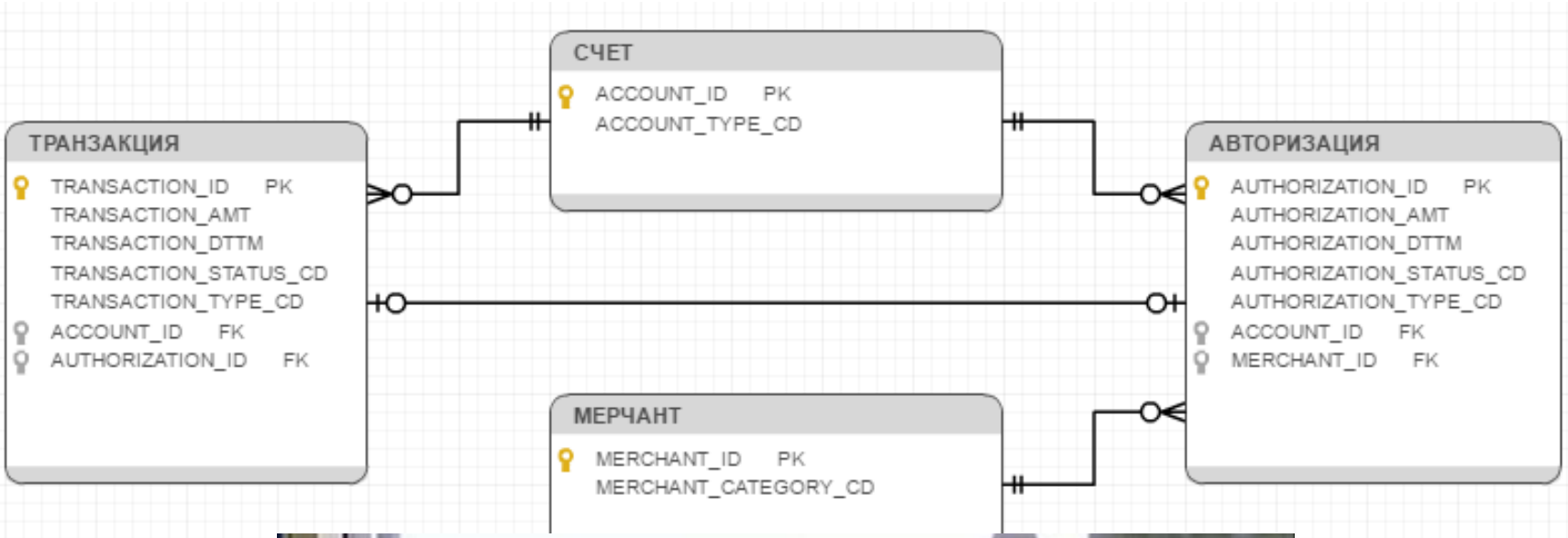
- Денормализованные витрины
- 3-я нормальная форма
- Data Vault

- Унификация алгоритмов загрузки
- Простота автоматизации проектирования
- Простота разработки процессов загрузки (вплоть до полной автоматизации)
- Сложность аналитических запросов
- Сложность первоначального проектирования системы

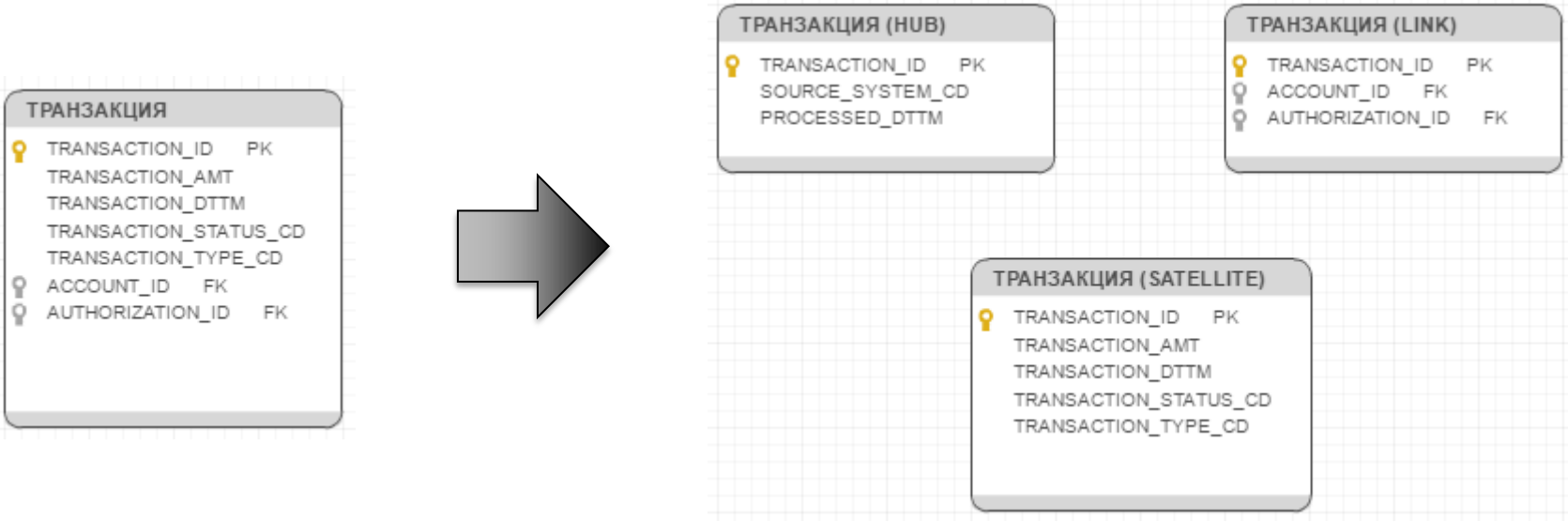
Денормализованные витрины



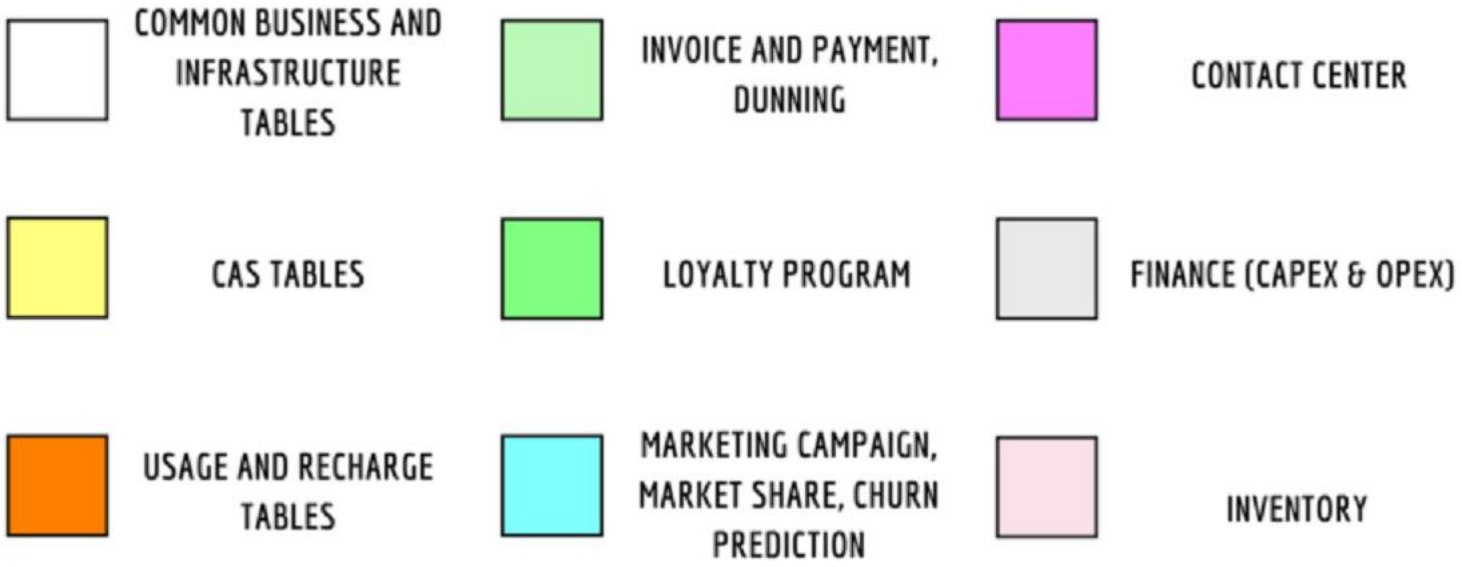
3-я нормальная форма



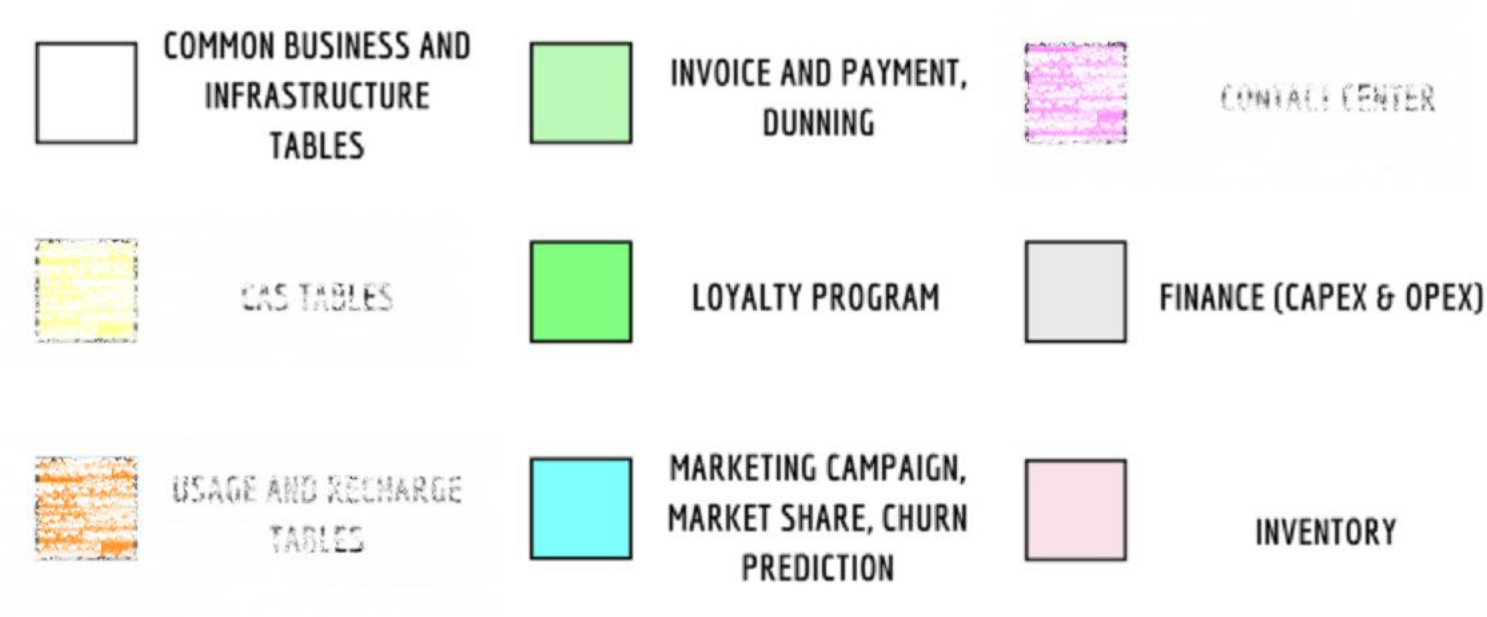
Data Vault


ТРАНЗАКЦИЯ
TRANSACTION_ID PK
TRANSACTION_AMT
TRANSACTION_DTTM
TRANSACTION_STATUS_CD
TRANSACTION_TYPE_CD
ACCOUNT_ID FK
AUTHORIZATION_ID FK
ТРАНЗАКЦИЯ (HUB)
TRANSACTION_ID PK
SOURCE_SYSTEM_CD
PROCESSED_DTTM
ТРАНЗАКЦИЯ (LINK)
TRANSACTION_ID PK
ACCOUNT_ID FK
AUTHORIZATION_ID FK
ТРАНЗАКЦИЯ (SATELLITE)
TRANSACTION_ID PK
TRANSACTION_AMT
TRANSACTION_DTTM
TRANSACTION_STATUS_CD
TRANSACTION_TYPE_CD

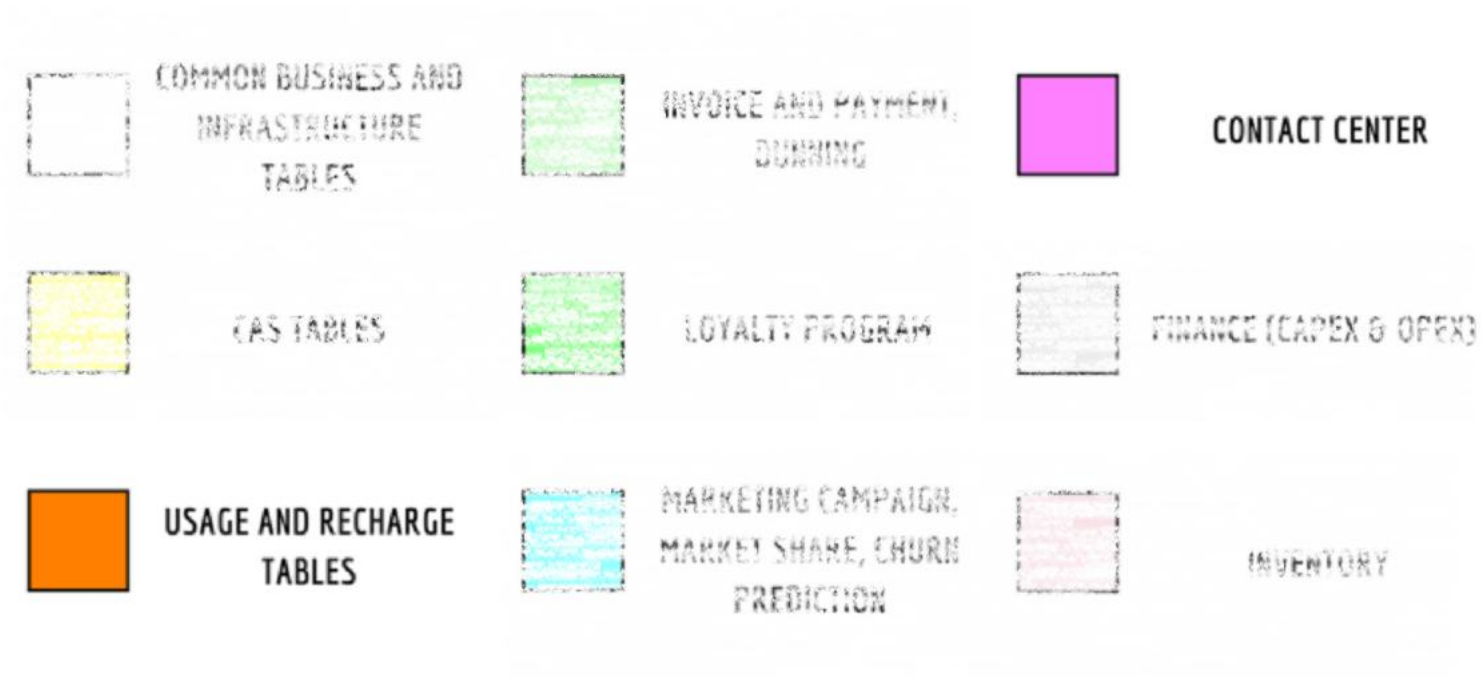
Пример. Модель телекома.



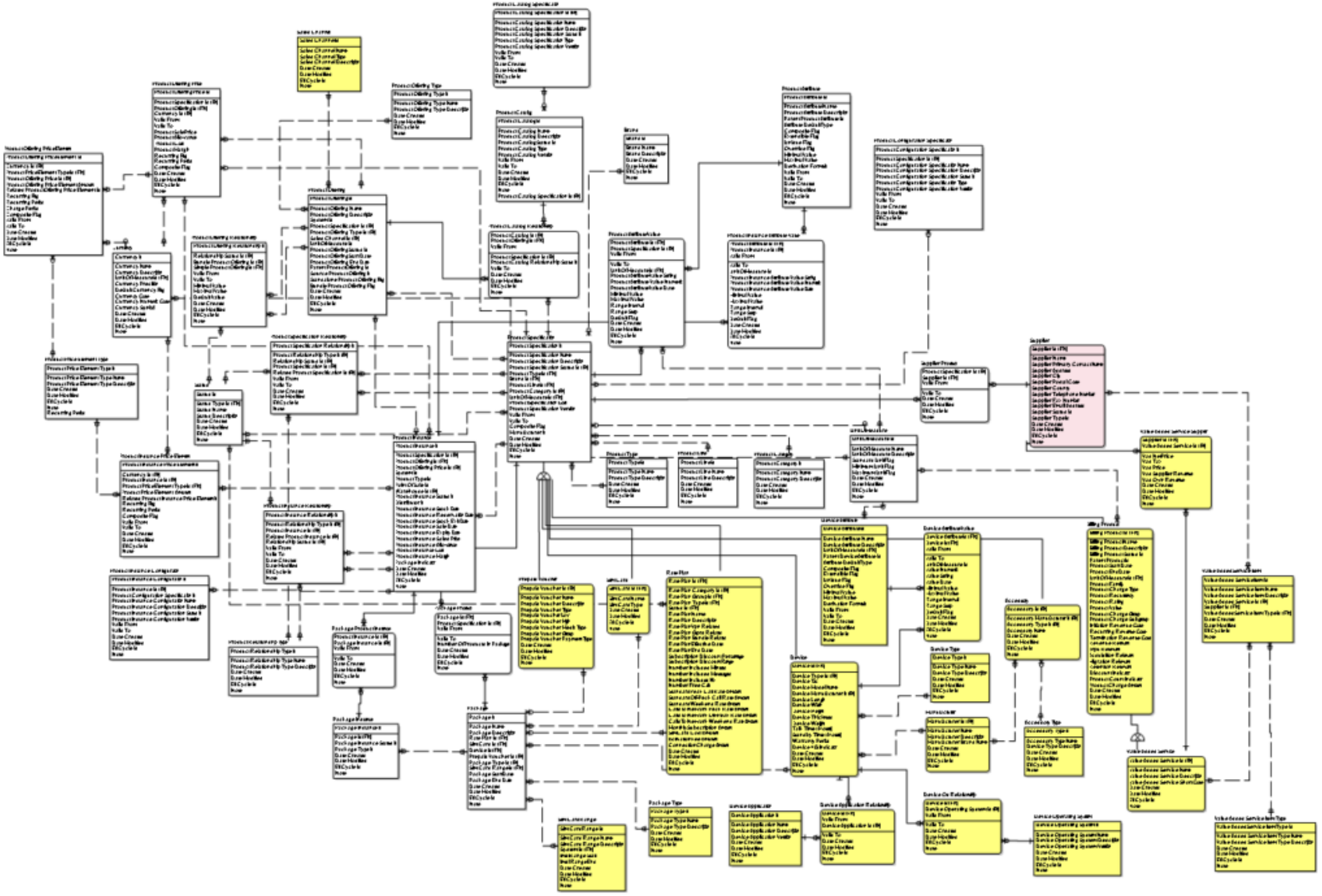
Пример. Модель телекома. sql



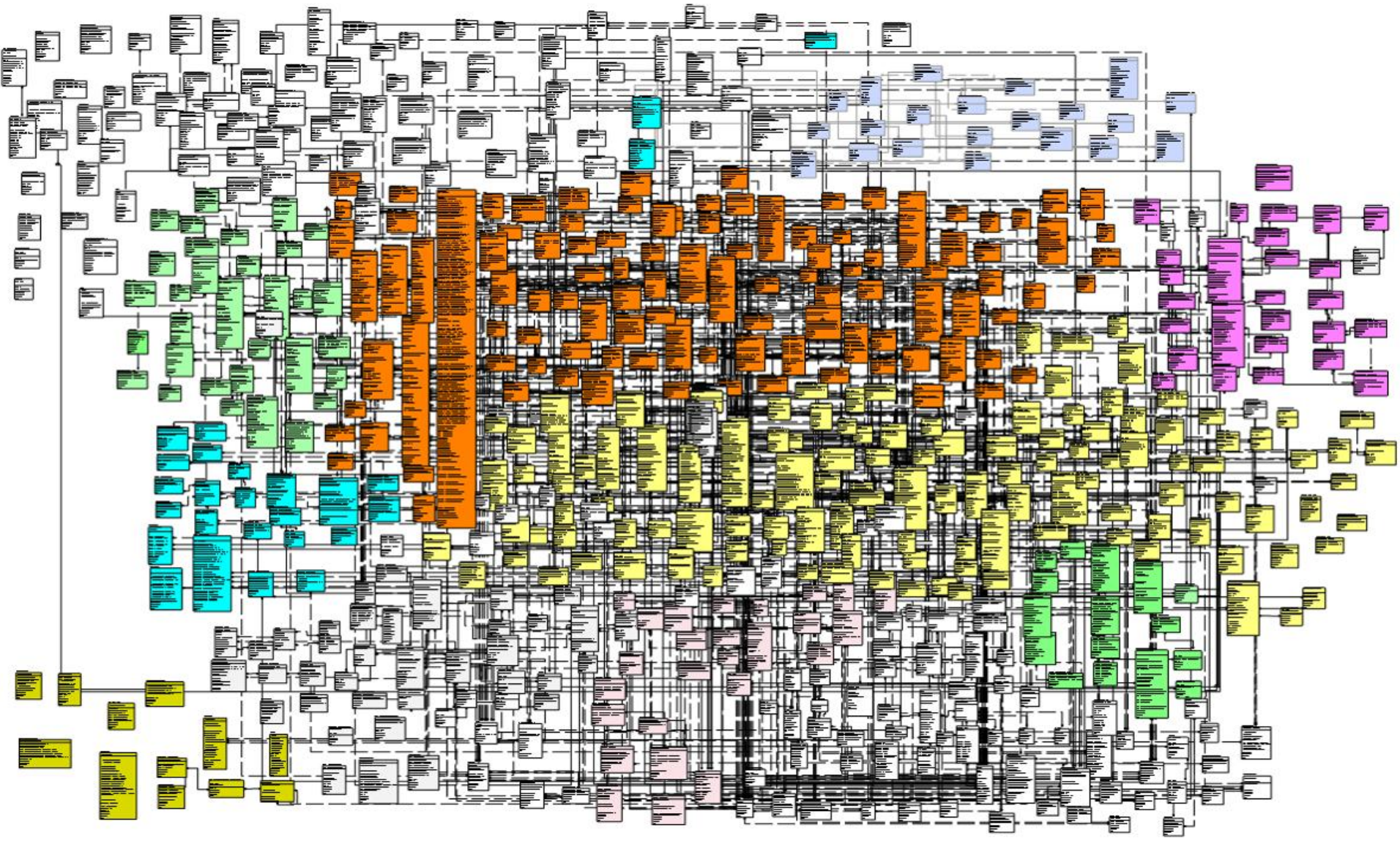
Пример. Модель телекома. sql VS nosql



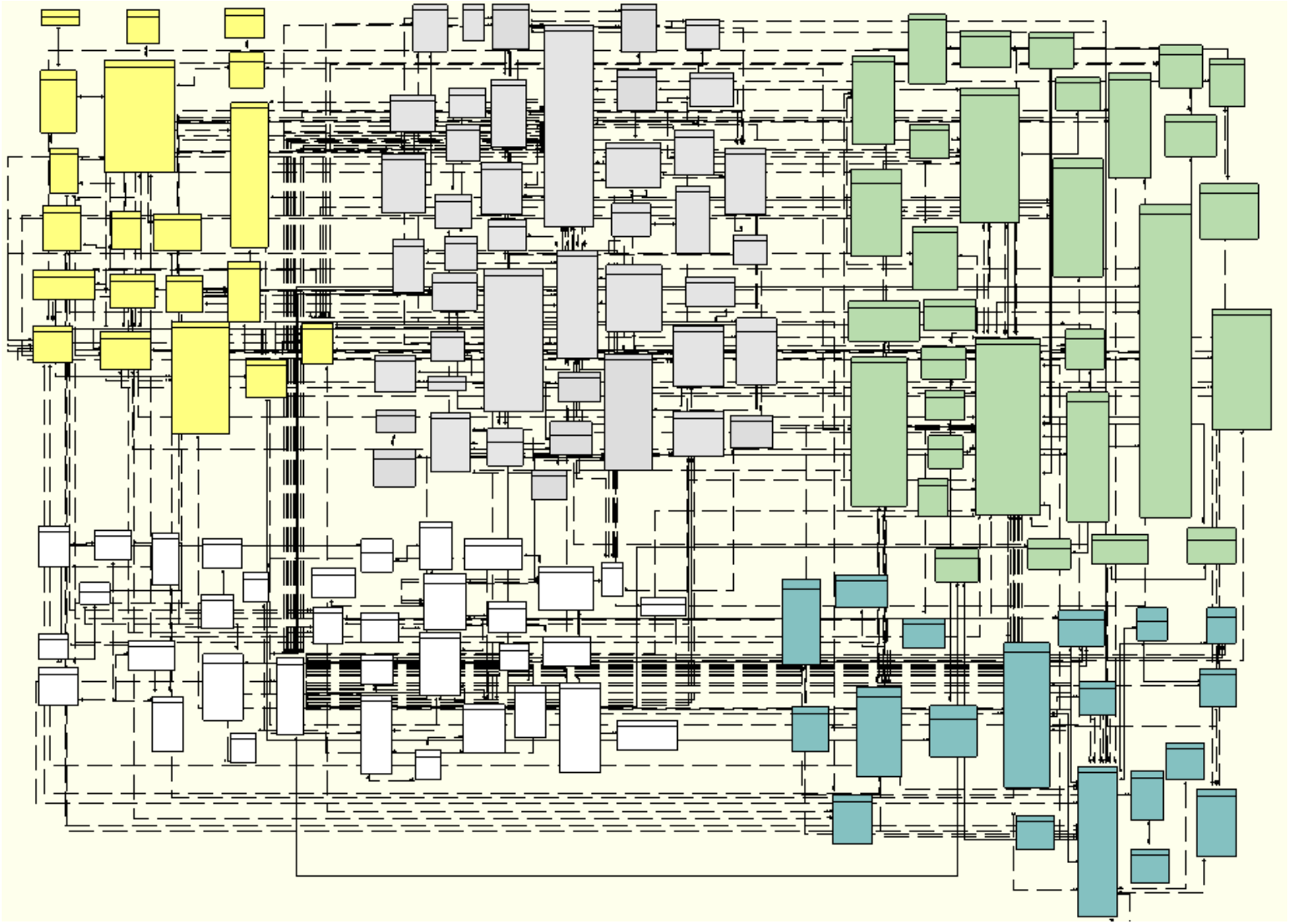
Пример. Модель телекома. Подобласть продукт



Пример. Модель телекома



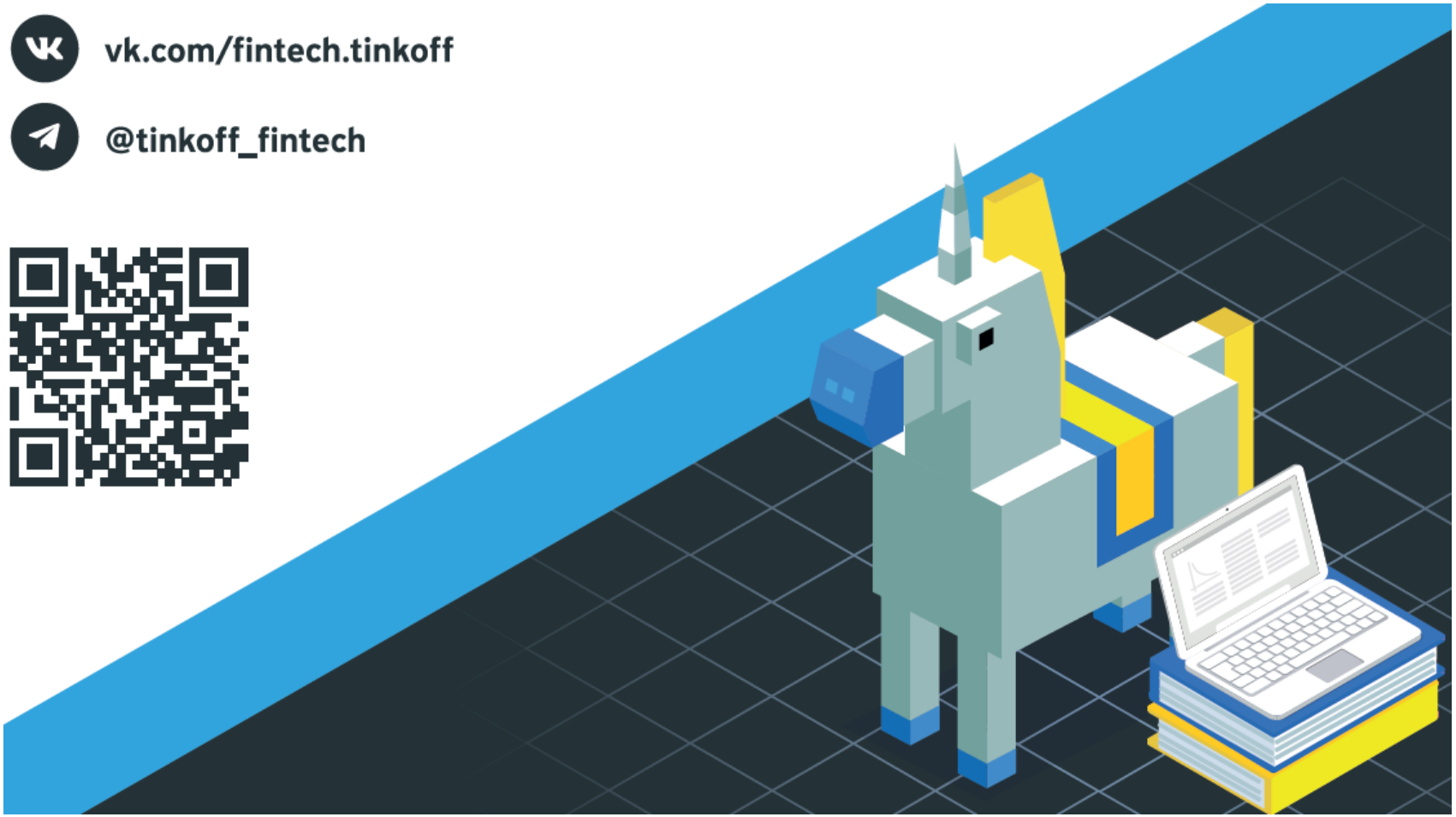
Пример. Модель страховой



noSQL

Что дальше?

https://fintech.tinkoff.ru



Тинькофф

Дальше действовать будем мы!

Tinkoff.ru