

# Модели данных и нормализация таблиц. Схема «звезда».

Урок 1



## Антон Иоффе

Senior Data Scientist, SAP

Последние 5 лет работаю в Data Science.  
Начинал как backend разработчик. Живу в Израиле.

- ✦ Система анонимизации видеофайлов, поиск поддельных чеков и счетов, автоматический аудит командировочных расходов и т.д.
- ✦ Запатентовано 7 различных алгоритмов связанных с машинным обучением








# План курса





## Что будет на уроке сегодня

-  Основные функции ETL-систем
-  Как работает ETL-система
-  Модели данных
-  Нормализация таблиц
-  Схема звезда



# Основные функции ETL-систем





## Что такое ETL

**ETL** – аббревиатура от Extract, Transform, Load. Это общий термин для процессов, которые происходят, когда данные переносят из нескольких систем в одно хранилище.

ETL – один из центральных процессов в системах хранения данных. Он включает в себя:

- **извлечение** данных из различных источников
- **трансформация** и очистка данных для приведения их к единообразию или в соответствие с бизнес-задачами.
- **загрузка** в хранилище данных



## Основные функции ETL

1

### Извлечение

Во время извлечения данных необработанные данные копируются или экспортируются из исходных местоположений в промежуточную область

2

### Трансформация

Данные преобразуются и консолидируются для предполагаемого аналитического использования

3

### Загрузка

На этом последнем шаге преобразованные данные перемещаются из промежуточной области в целевое хранилище данных



Вопрос

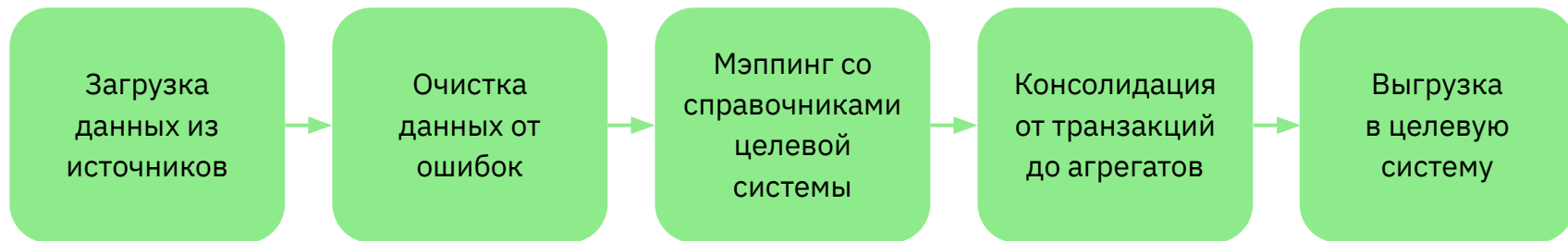
Зачем нужна ETL-система?







## Как работает ETL-система





## Лучшие практики ETL

### **Планируйте очистку данных**

Всегда планируйте что-то чистить, потому что главная причина создания хранилища данных — предлагать более чистые и надежные данные

### **Не пытайтесь очистить все данные**

Каждая организация хотела бы, чтобы все данные были чистыми, но большинство из них не готовы платить за ожидание или не готовы ждать. Очистка всего этого займет слишком много времени, поэтому лучше не пытаться очистить все данные

### **Определите стоимость очистки**

Перед очисткой всех «грязных» данных важно определить стоимость очистки для каждого «грязного» элемента данных.

### **Имейте вспомогательные индексы**

Чтобы сократить расходы на хранение, храните обобщенные данные на дисках. Также требуется компромисс между объемом хранимых данных и их подробным использованием



# Модели данных и нормализация таблиц





# Модели данных

1

## Иерархическая

представляет собой совокупность элементов, расположенных в порядке их подчинения от общего к частному и образующих перевернутое по структуре дерево (граф).

2

## Реляционная

объекты и связи между ними представляет в виде таблиц, при этом связи тоже рассматриваются как объекты. Все строки, составляющие таблицу в реляционной базе данных, должны иметь первичный ключ

3

## Сетевая

каждый элемент может быть связан с любым другим элементом



## Нормализация таблиц

**Нормализация** представляет собой процесс реорганизации данных путем ликвидации избыточности данных и иных противоречий с целью приведения таблиц к виду, позволяющему осуществлять непротиворечивое и корректное редактирование данных.

Как правило, нормализация преимущественно применяется при восходящем подходе проектировании базы данных, то есть когда мы все атрибуты, которые надо сохранить в бд, группируем по сущностям, для которых затем создаются таблицы.

**В ненормализованной форме** таблица может хранить информацию о двух и более сущностях. Также она может содержать повторяющиеся столбцы. Также столбцы могут хранить повторяющиеся значения. В нормализованной же форме каждая таблица хранит информацию только об одной сущности.

Нормализация предполагает применение нормальных форм к структуре данных. Существуют 7 нормальных форм. Каждая нормальная форма (за исключением первой) подразумевает, что к данным уже была применена предыдущая нормальная форма.



## Первая нормальная форма

Марка	Модель
Nissan	GT-R
Porsche	911, Cayenne, Taycan

Марка	Модель
Nissan	GT-R
Porsche	911
Porsche	Cayenne
Porsche	Taycan



## Вторая нормальная форма

Модель	Марка	Цена	Скидка
GT-R	Nissan	7500000	10%
911	Porsche	15000000	5%
Cayenne	Porsche	12000000	5%
Taуcan	Porsche	9000000	5%

Модель	Марка	Цена
GT-R	Nissan	7500000
911	Porsche	15000000
Cayenne	Porsche	12000000
Taуcan	Porsche	9000000

Марка	Скидка
Nissan	10%
Porsche	5%



## Третья нормальная форма

Марка	Автосалон	Телефон
Nissan	Люкс-авто	745-32-05
Porsche	Люкс-авто	745-32-05
Lada	Жигуль-топ	733-11-14

Автосалон	Телефон
Люкс-авто	745-32-05
Жигуль-топ	733-11-14

Марка	Автосалон
Nissan	Люкс-авто
Porsche	Люкс-авто
Lada	Жигуль-топ





## Нормальная форма Бойса-Кодда

Номер автомобиля	Время начала	Время окончания	Тариф
1	09:30	10:30	Эконом
1	11:00	12:00	Эконом
1	14:00	15:30	Стандарт
2	10:00	12:00	Премиум-В
2	12:00	14:00	Премиум-В
2	15:00	18:00	Премиум-А

Тариф	Номер автомобиля	Имеет льготы
Эконом	1	Да
Стандарт	1	Нет
Премиум-А	2	Да
Премиум-В	2	Нет

Тариф	Время начала	Время окончания
Эконом	09:30	10:30
Эконом	11:00	12:00
Стандарт	14:00	15:30
Премиум-В	10:00	12:00
Премиум-В	12:00	14:00
Премиум-А	15:00	18:00



## Другие нормальные формы

### Четвертая нормальная форма

Отношение находится в 4НФ, если оно находится в НФБК и все нетривиальные многозначные зависимости фактически являются функциональными зависимостями от ее потенциальных ключей.

### Пятая нормальная форма

Отношения находятся в 5НФ, если оно находится в 4НФ и отсутствуют сложные зависимые соединения между атрибутами

### Доменно-ключевая нормальная форма

Переменная отношения находится в ДКНФ тогда и только тогда, когда каждое наложенное на неё ограничение является логическим следствием ограничений доменов и ограничений ключей, наложенных на данную переменную отношения



## Шестая нормальная форма

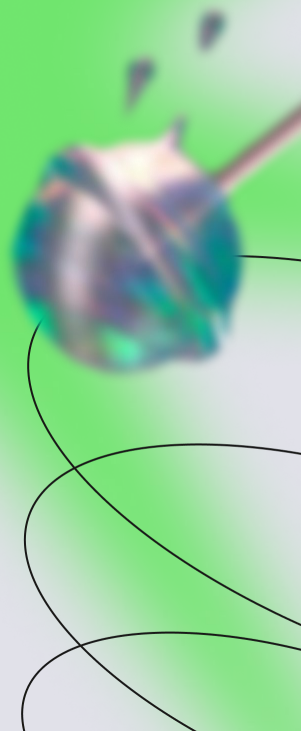
№	Время	Должность	Телефон
<b>3322</b>	01-03-2020:11-04-2022	аналитик	+79374422755
<b>3124</b>	31-07-2019:14-03-2022	программист	+79374443111
<b>3322</b>	12-04-2022:21-09-2022	программист	+79374422755

№	Время	Должность
<b>3322</b>	01-03-2020:11-04-2022	аналитик
<b>3124</b>	31-07-2019:14-03-2022	программист

№	Время	Телефон
<b>3322</b>	01-03-2020:11-04-2022	+79374422755
<b>3124</b>	31-07-2019:14-03-2022	+79374443111



Хранилища данных





# Хранилища данных

## Концепция хранилища данных

Основная концепция хранилища данных состоит в том, чтобы упростить для компании единую версию правды для принятия решений и прогнозирования.

**Многомерная схема** специально разработана для моделирования систем хранилищ данных.

Три основных типа многомерных схем, каждая из которых имеет свои уникальные преимущества.

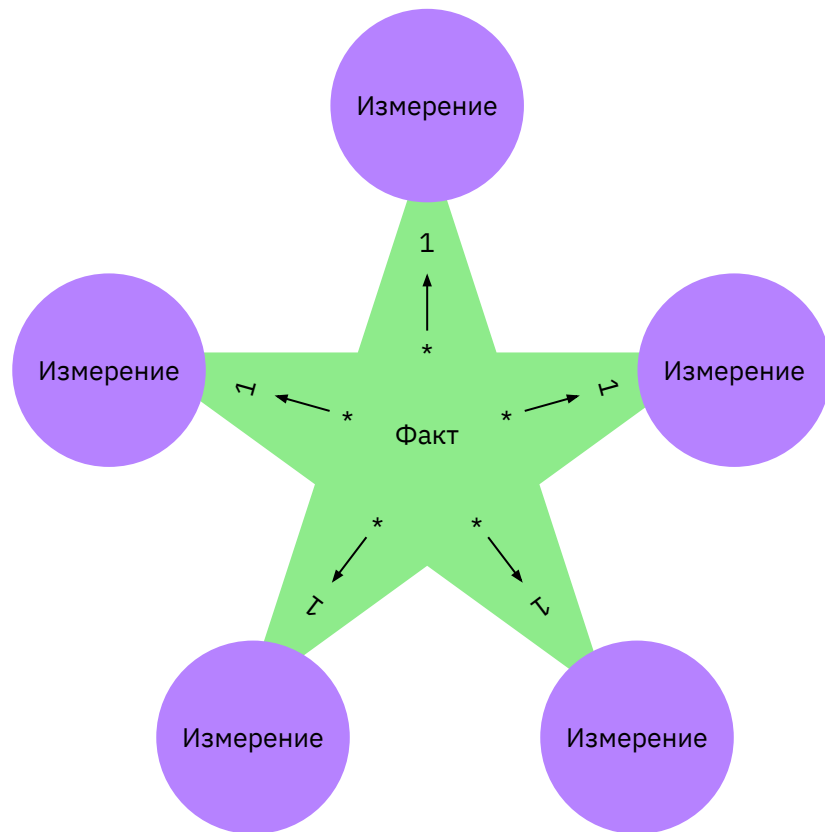
- Схема звезда
- Схема снежинка
- Схема галактика



## Схема звезда

Данная схема основана на построении нормализованных таблиц из которых, в случае чего, могут быть получены денормализованные таблицы. Основные характеристики:

- Каждое измерение в звездообразной схеме представлено единственной одномерной таблицей.
- Таблица измерений должна содержать набор атрибутов.
- Таблица измерений присоединяется к таблице фактов с помощью внешнего ключа
- Таблица измерений не соединены друг с другом
- Таблица фактов будет содержать ключ и меру
- Схема звезда проста для понимания и обеспечивает оптимальное использование диска.
- Таблицы измерений **не нормализованы**.
- Схема широко поддерживается BI Tools





**Спасибо за внимание**

