SECR

Software Engineering Conference Russia 2018

October 12-13
Moscow

Аналитика на 💯 млн данных.

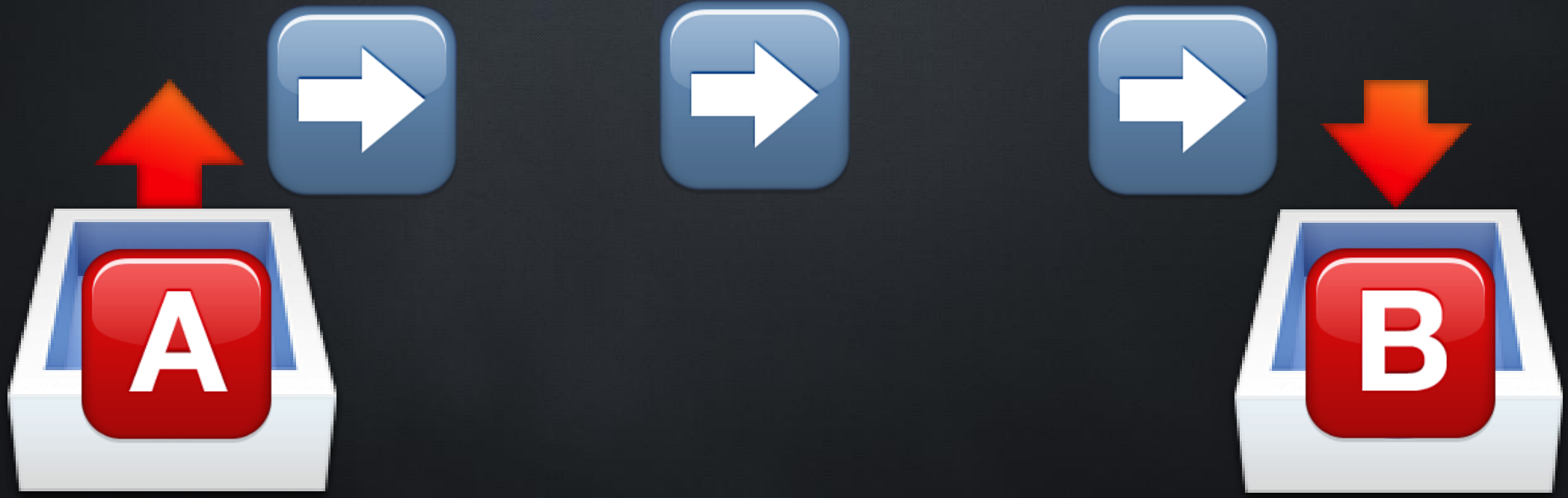
Краткий ликбез для системных интеграторов

Бунто Татьяна

H F Labs
data quality
customer data integration

human factor laboratories
hflabs.ru

О чем речь?
Какая аналитика? Какие данные?



Система-источник данных

Система-приемник

О чем речь?
Какая аналитика? Какие данные?

Система-источник данных

Реляционная БД

Система-приемник

Реляционная БД

При чем здесь системные интеграторы?

Система-источник данных

ETL-процесс или шина

выгрузка, преобразование и загрузка

Система-приемник

При чем здесь системные интеграторы?

Проводим аналитику данных на каждом этапе:

1. Что хранится в системе-источнике

При чем здесь системные интеграторы?

Проводим аналитику данных на каждом этапе:

1. Что хранится в системе-источнике

 1 человек и  20 000 его лицевых счетов  как храним?

При чем здесь системные интеграторы?

Проводим аналитику данных на каждом этапе:

1. Что хранится в системе-источнике

 1 человек и  20 000 его лицевых счетов  как храним?

2. Обработка ETL-процессом или шиной

При чем здесь системные интеграторы?

Проводим аналитику данных на каждом этапе:

1. Что хранится в системе-источнике

 1 человек и  20 000 его лицевых счетов  как храним?

2. Обработка ETL-процессом или шиной

 было 100 млн клиентов → после загрузки осталось 2 млн 

При чем здесь системные интеграторы?

Проводим аналитику данных на каждом этапе:

1. Что хранится в системе-источнике

 1 человек и  20 000 его лицевых счетов  как храним?

2. Обработка ETL-процессом или шиной

 было 100 млн клиентов → после загрузки осталось 2 млн 

3. Что сохраняется в системе-приемнике

При чем здесь системные интеграторы?

Проводим аналитику данных на каждом этапе:

1. Что хранится в системе-источнике

 1 человек и  20 000 его лицевых счетов  как храним?

2. Обработка ETL-процессом или шиной

 было 100 млн клиентов → после загрузки осталось 2 млн 

3. Что сохраняется в системе-приемнике

были  домашние и  мобильные  стали только домашние 

Чем я занимаюсь?

Чем я занимаюсь?

Кредитная карта
ФИО: Щукин Сергей Владимирович
ДР: 16.03.1982
Телефон:+7 123 456 78 90
Паспорт: 12 34 123456

Чем я занимаюсь?

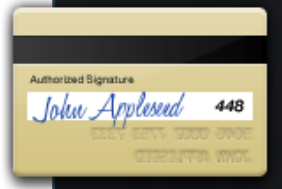
Кредитная карта
ФИО: Щукин Сергей Владимирович
ДР: 16.03.1982
Телефон:+7 123 456 78 90
Паспорт: 12 34 123456

Дебетовая карта
ФИО: SERGEY SCHOUKIN
ДР: 16.03.1982

Чем я занимаюсь?

Кредитная карта
ФИО: Щукин Сергей Владимирович
ДР: 16.03.1982
Телефон:+7 123 456 78 90
Паспорт: 12 34 123456

Дебетовая карта
ФИО: SERGEY SCHOUKIN
ДР: 16.03.1982

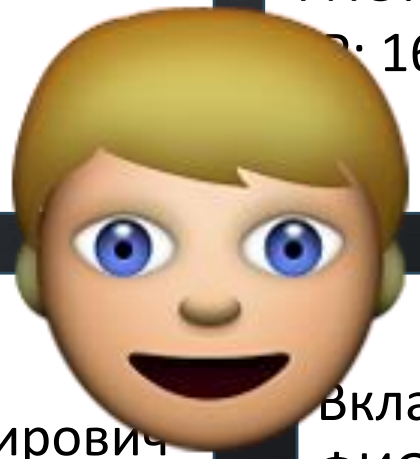


Ипотечный договор
ФИО: Щукин Сергей Владимирович
Телефон: +7 234 567 89 01
Паспорт: 12 34 123456
Адрес: 127642, Москва, Мира 2-164

Чем я занимаюсь?

Кредитная карта
ФИО: Щукин Сергей Владимирович
ДР: 16.03.1982
Телефон: +7 123 456 78 90
Паспорт: 12 34 123456

Дебетовая карта
ФИО: SERGEY SCHOUKIN
ДР: 16.03.1982



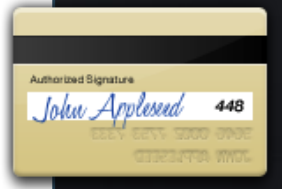
Ипотечный договор
ФИО: Щукин Сергей Владимирович
Телефон: +7 234 567 89 01
Паспорт: 12 34 123456
Адрес: 127642, Москва, Мира 2-164

Вклад
ФИО: Щукин Сергей Владимирович
Телефон: +7 234 567 89 01
ДР: 16.03.1982

Чем я занимаюсь?

Кредитная карта
ФИО: Щукин Сергей Владимирович
ДР: 16.03.1982
Телефон: +
Паспорт: 1

Дебетовая карта
ФИО: SERGEY SCHOUKIN



«Единый клиент»

ФИО: Щукин Сергей Владимирович
ДР: 16.03.1982
Телефон: +7 123 456 78 90
+7 234 567 89 01
Паспорт: 12 34 123456
Адрес: 127642, г. Москва, пр. Мира, д. 2, кв. 164

Ипотечный
ФИО: Щук
Телефон: +
Паспорт: 7
Адрес: 127

димирович

Интеграции «Единого клиента»

1. Онлайн-поток
2. Пакетный инкремент (буферные таблицы, представления БД)

Зачем нужна эта аналитика?
У меня уже есть все необходимое
Бери и интегрируйся

Документация на руках

Зачем нужна эта аналитика?
У меня уже есть все необходимое
Бери и интегрируйся

 Документация на руках

 Бизнес-процессы изучены

Зачем нужна эта аналитика?
У меня уже есть все необходимое
Бери и интегрируйся

 Документация на руках

 Бизнес-процессы изучены

 Архитектор все рассказал про интеграции

Зачем нужна эта аналитика?
У меня уже есть все необходимое
Бери и интегрируйся


Документация на руках


Бизнес-процессы изучены


Архитектор все рассказал про интеграции


Мэппинг передачи данных согласовали

Зачем нужна эта аналитика?
У меня уже есть все необходимое
Бери и интегрируйся

⛔ Документация на руках

 Документация ≠ реальность

✅ Бизнес-процессы изучены

✅ Архитектор все рассказал про интеграции

✅ Мэппинг передачи данных согласовали

Зачем нужна эта аналитика?
У меня уже есть все необходимое
Бери и интегрируйся

⛔ Документация на руках

⛔ Бизнес-процессы изучены

✅ Архитектор все рассказал про интеграции

✅ Мэппинг передачи данных согласовали

Документация ≠ реальность

Человеческий фактор никто не отменял

Зачем нужна эта аналитика?
У меня уже есть все необходимое
Бери и интегрируйся

🚫 Документация на руках

💣 Документация ≠ реальность

🚫 Бизнес-процессы изучены

😴 Человеческий фактор никто не отменял

🚫 Архитектор все рассказал про интеграции

👣 Система одна, архитекторы меняются

✅ Мэппинг передачи данных согласовали

Зачем нужна эта аналитика?
У меня уже есть все необходимое
Бери и интегрируйся

- Документация на руках
- Бизнес-процессы изучены
- Архитектор все рассказал про интеграции
- Мэппинг передачи данных согласовали

Документация ≠ реальность

Человеческий фактор никто не отменял

Система одна, архитекторы меняются

Уверены, что ничего лишнего не загрузите?

Зачем нужна эта аналитика?
У меня уже есть все необходимое
Бери и интегрируйся

- Докумен
- ...л
- ...ются
- Мэппинг переда... ...ны, что ничего лишнего не загрузите?

ВСЕ ВРУТ!!!©

Цель, средства и планирование

Из жизни одного проекта

Аналитика только на последнем этапе:

1. Перегрузка данных — 20 дней
2. Пересогласование и переделывание модели — 15 дней
3. Лишние данные и потеря необходимых — 3 перегрузки

Цель, средства и планирование

Из жизни одного проекта

Аналитика только на последнем этапе:

1. Перегрузка данных — 20 дней
2. Пересогласование и переделывание модели — 15 дней
3. Лишние данные и потеря необходимых — 3 перегрузки

Аналитика и выводы ~ 10 дней

(3 дня система-источник, 3 дня ETL, 3 дня система-приемник, 1 день на отчеты и переписки)

100 млн, 10 млн, 1 млн данных
Есть ли разница?

100 млн, 10 млн, 1 млн данных
Есть ли разница?

1. Цена ошибки потери информации

100 млн. → миллионом больше, миллионом меньше → не так очевидно

100 млн, 10 млн, 1 млн данных
Есть ли разница?

1. Цена ошибки потери информации

   100 млн. → миллионом больше, миллионом меньше → не так очевидно

2. Время перегрузки и точки анализа информации

   перегрузка 100 млн. → 20 дней

   перегрузка 1 млн. → 1 день

Как анализируем?

Простые SQL- запросы, шаблон готовим в Экселе

Как анализируем?

Простые SQL- запросы, шаблон готовим в Экселе

Скрипты заполненности

	A	B	C	C
	Поле	Таблица	Формула	Результат
1			select	select
2	TOTAL	BUFFER_PH	count(*) TOTAL	count(*) TOTAL
3	LAST_NM	BUFFER_PH	=СЦЕПИТЬ(", count(";A3;") "; A3)	, count(LAST_NM) LAST_NM
4	FIRST_NM	BUFFER_PH	=СЦЕПИТЬ(", count(";A4;") "; A4)	, count(FIRST_NM) FIRST_NM
…	…	…	…	…
8	DELETED	BUFFER_PH	=СЦЕПИТЬ(", count(";A8;") "; A8)	, count(DELETED) DELETED
9			from BUFFER_PH;	from BUFFER_PH;

Все числа и примеры являются реальными, любое совпадение с существующими данными и системами не случайно!

Что анализируем?

1. Заполненность или наличие null-значений
2. Длину полей в БД
3. Распределение значений по длине
4. Распространенность или популярность
5. Наличие справочников и классификаторов
6. Консистентность
7. Адекватность данных

1. Заполненность или null-значения

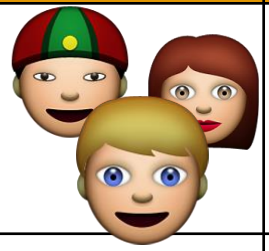
1. Сколько всего строк в таблице

```
Select count(*) from <table>;
```

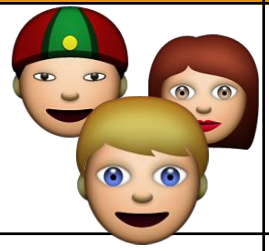
2. Сколько заполнено в каждом столбце

```
Select <column_name>, count(*) as <column_name> cnt from <table>
 where <column_name> is not null;
```

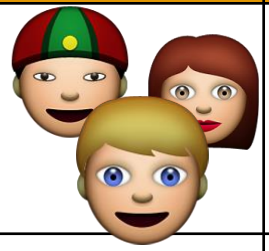
1. Заполненность или null-значения

Таблица ФЛ	Выгрузка 1	Выгрузка 2	Дельта	Вывод
Всего 	99 966 324	94 847 160	-5 119 164	В декабре загрузили лишние технические записи

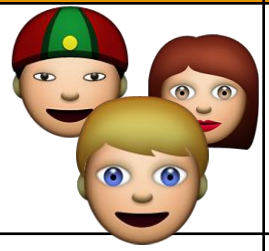
1. Заполненность или null-значения

Таблица ФЛ	Выгрузка 1	Выгрузка 2	Дельта	Вывод
Всего 	99 966 324	94 847 160	-5 119 164	В декабре загрузили лишние технические записи
ДР	0	77 046 780	77 046 780	Надо разобраться

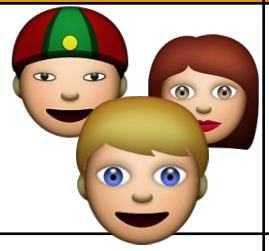
1. Заполненность или null-значения

Таблица ФЛ	Выгрузка 1	Выгрузка 2	Дельта	Вывод
Всего 	99 966 324	94 847 160	-5 119 164	В декабре загрузили лишние технические записи
ДР	0	77 046 780	77 046 780	Надо разобраться
ИНН	65 136	74 591	9 455	Посмотреть детальнее

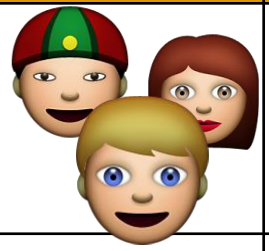
1. Заполненность или null-значения

Таблица ФЛ	Выгрузка 1	Выгрузка 2	Дельта	Вывод
Всего 	99 966 324	94 847 160	-5 119 164	В декабре загрузили лишние технические записи
ДР	0	77 046 780	77 046 780	Надо разобраться
ИНН	65 136	74 591	9 455	Посмотреть детальнее
СНИЛС	0	0	0	Зачем он в модели?

1. Заполненность или null-значения

Таблица ФЛ	Выгрузка 1	Выгрузка 2	Дельта	Вывод
Всего 	99 966 324	94 847 160	-5 119 164	В декабре загрузили лишние технические записи
ДР	0	77 046 780	77 046 780	Надо разобраться
ИНН	65 136	74 591	9 455	Посмотреть детальнее
СНИЛС	0	0	0	Зачем он в модели?
Язык	4 156	3 288 872	3 284 716	Потеряли в декабре

1. Заполненность или null-значения

Таблица ФЛ	Выгрузка 1	Выгрузка 2	Дельта	Вывод
Всего 	99 966 324	94 847 160	-5 119 164	В декабре загрузили лишние технические записи
ДР	0	77 046 780	77 046 780	Надо разобраться
ИНН	65 136	74 591	9 455	Посмотреть детальнее
СНИЛС	0	0	0	Зачем он в модели?
Язык	4 156	3 288 872	3 284 716	Потеряли в декабре
Флаг удаления	0	0	0	Клиентов не удаляют?!
…	…	…	…	…

1. Заполненность или null-значения

Адреса	Выгрузка 1	Выгрузка 2	Дельта	Вывод
Всего 	254 803 976	248 453 078	-6 350 898	В декабре прогрузили технические записи

1. Заполненность или null-значения

Адреса	Выгрузка 1	Выгрузка 2	Дельта	Вывод
Всего 	254 803 976	248 453 078	-6 350 898	В декабре прогрузили технические записи
Страна	229 256 090	220 760 236	-8 495 854	!!! Сравниваем с общей заполненностью и идем к бизнесу уточнять, как будут использованы эти данные!
Индекс	46 834 777	50 954 835	4 120 058	
Город	6 474 841	7 497 034	1 022 193	
Улица	894 040	834 073	-59 967	
Дом	20 903	22 753	1 850	
…	…	…	…	…

1. Заполненность или null-значения

1. Если во всей выгрузке какое-то поле Null — проверяем, точно ли этих данных нет в исходной системе, может их потеряли при выгрузке
2. Слабая заполненность поля. Интересуемся «почему так мало» и нужны ли эти данные в вашей системе?
3. Если поле — справочник, то оно, как правило, не бывает Null. Справочники проверяем отдельно!

2. Длина данных — очень короткие или максимальные по длине

1. Короткие по длине

   Select * from <table> where length (column_name)<3;

2. Максимальная длина колонки (допустим 256 символов)

   Select * from cdi_buffer_physical where length (column_name)=256;

2. Длина данных — очень короткие или максимальные по длине

1. Короткие по длине. Большая вероятность мусорных значений или кодировок

2. Длина данных — очень короткие или максимальные по длине

1. Короткие по длине. Большая вероятность мусорных значений или кодировок

Фамилия



2. Максимальные по длине. Заполненность колонки впритык — частый маркер того, что данные не помещались в колонку по длине и были обрезаны

Контактный фиксированный телеф

3. Распределение значений по длине

Если есть много данных, одинаковых по длине, но не вписывающихся в статистику, есть вероятность, что в исходную систему, когда-то смигрировали обрезанные данные

```
Select length (column_name), count(column_name) from <table>

group by length (column_name);
```

3. Распределение значений по длине

Если есть много данных, одинаковых по длине, но не вписывающихся в статистику, есть вероятность, что в исходную систему, когда-то смигрировали обрезанные данные

```
Select length (column_name), count(column_name) from <table>

group by length (column_name);
```

Длина	Количество
6	20
4	2
10	3

Россия

Герман

Велико

4. Распространенность или популярность

Смотрим топ-распределений строковых значений и все распределения справочных значений

1. Для справочников

   select <column_name>,count(*) cnt from <table>

   group by <column_name> order by 2 desc;

2. Для строковых значений ограничиваем число результатов в выводе

   select * from (select <column_name> ,count(*) cnt from <table>

   group by <column_name> order by 2 desc) where rownum<=100;

4. Распространенность или популярность

Имя	Количество
-	541 727
*	425 965
Тест	333 789
Александр	192 834

4. Распространенность или популярность

Имя	Количество
-	541 727
*	425 965
Тест	333 789
Александр	192 834

Дата рождения	Количество
	9 314 770
01.01.1980	164 866
01.01.1900	117 078
15.09.2015	53 702

4. Распространенность или популярность

Имя	Количество
-	541 727
*	425 965
Тест	333 789
Александр	192 834

Дата рождения	Количество
	9 314 770
01.01.1980	164 866
01.01.1900	117 078
15.09.2015	53 702

Страна выдачи документа	Количество
Россия	55 888 454
	7 169 636
Россия.	6 529 147
Антарктида	216 475

4. Распространенность или популярность

Имя	Количество
-	541 727
*	425 965
Тест	333 789
Александр	192 834

Дата рождения	Количество
	9 314 770
01.01.1980	164 866
01.01.1900	117 078
15.09.2015	53 702

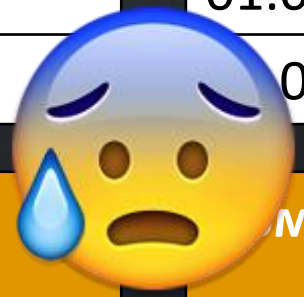
Страна выдачи документа	Количество
Россия	55 888 454
	7 169 636
Россия.	6 529 147
Антарктида	216 475

Номер документа	Количество
0	112 559
	70 642
1	67 059
О	29 807

4. Распространенность или популярность

Имя	Количество
-	541 727
*	425 965
Тест	333 789
Александр	192 834

Дата рождения	Количество
	9 314 770
01.01.1980	164 866
01.01.1900	117 078
09.2015	53 702



Страна выдачи документа	Количество
Россия	55 888 454
	7 169 636
Россия.	6 529 147
Антарктида	216 475

мер документа	Количество
0	112 559
	70 642
1	67 059
О	29 807

4. Распространенность или популярность

Самые популярные значения могут быть:

1. Просто самыми популярными. Имя «Татьяна» или отчество «Владимирович». Они не должны выбиваться из статистики. Татьяна не может быть в 10 раз популярнее Натальи, а Исмаил популярнее Александра
2. Мусорными значениями
3. Автозаполнением на стороне исходной системы

5. Справочники и классификаторы

Отбираем поля, которые в реальной жизни обычно являются справочниками и проверяем, являются ли эти поля справочниками в выгрузке

5. Справочники и классификаторы

Отбираем поля, которые в реальной жизни обычно являются справочниками и проверяем, являются ли эти поля справочниками в выгрузке

Место рождения	Количество
таджикистан	467 599
Таджикистан	410 484
Россия	292 585
ТАДЖИКИСТАН	234 465
россия	158 163
РОССИЯ	76 367

Проблемы неунифицированных значений из-за:

- опечаток
- пробелов
- разного регистра
- прочего фольклора

5. Справочники и классификаторы

Проверяем, какие поля внезапно являются справочниками

Должность. Выбор в личном кабинете:


Директор


Бухгалтер


Специалист


Секретарь


Системный администратор

5. Справочники и классификаторы

Отсутствие ожидаемых справочных значений

Пол:

1. Женский

❓

2. Не определен

5. Справочники и классификаторы

Отсутствие ожидаемых справочных значений

Пол:

1. Женский

❓

2. Не определен

Тип телефона:

1. Домашний

❓

2. Не определен

5. Справочники и классификаторы

Отсутствие ожидаемых справочных значений

Пол:

1. Женский

❓

2. Не определен

Язык:

1. Русский

Тип телефона:

1. Домашний

❓

2. Не определен

6. Консистентность и кросс-сверки

Все данные, которые должны быть связаны, связаны на самом деле

```
select count(*) from ((select <ID1> from <table1>)

minus

(select <ID2> from <table2>));
```

Уникальные данные — уникальны

```
select count(*), count(distinct doc_rk), PARTY_TYPE from BUFFER_DOC

where PT_SOURCE_SYSTEM_CD='A'

group by PARTY_TYPE;
```

6. Консистентность и кросс-сверки

Телефоны привязаны к несуществующим клиентам

6. Консистентность и кросс-сверки

Телефоны привязаны к несуществующим клиентам

Повторяются идентификаторы — у вас это первичные ключи

7. Адекватность данных

Полная свобода творчества

7. Адекватность данных

Полная свобода творчества

1 клиент — 50 000 телефонов. Это нормально?

7. Адекватность данных

Полная свобода творчества

1 клиент — 50 000 телефонов. Это нормально?

ИНН вида 79853617764, 89109462345, 4956780966.

Да это же телефоны?!

7. Адекватность данных

Полная свобода творчества

1 клиент — 50 000 телефонов. Это нормально?

ИНН вида 79853617764, 89109462345, 4956780966.

Да это же телефоны?!

Гранулярный адрес — Москва | Турчанинов | 6с2

Полный адрес — Санкт-Петербург, Невский проспект, 88

На что еще посмотреть?

Наличие латинских символов там, где их не должно быть

```
select <column_name> from <table>

where regexp_like(<column_name>, '[A-Z]', 'i');
```

Буквы и символы в строковых полях, предназначенных для цифр

```
select <column_name> from <table>

where regexp_like(<column_name>, ‘[^0-9]');
```

Вам помогут регулярные выражения!!!

Что делать с результатами?

1. Рассказываем системе-источнику о возможных улучшениях и наличии проблем

Что делать с результатами?

1. Рассказываем системе-источнику о возможных улучшениях и наличии проблем

2. Оптимизируем свою модель данных

Что делать с результатами?

1. Рассказываем системе-источнику о возможных улучшениях и наличии проблем

2. Оптимизируем свою модель данных

3. Выявляем и устраняем проблемы до ввода системы в эксплуатацию

Одни только цифры при аналитике ничего не значат и не несут смысловой нагрузки!

❗ Одни только цифры при аналитике ничего не значат и не несут смысловой нагрузки!

❗ Не забывайте смотреть на данные!

Одни только цифры при аналитике ничего не значат и не несут смысловой нагрузки!

Не забывайте смотреть на данные!

Предоставляйте заказчику выводы, а не только цифры!

Бунто Татьяна

tbunto89@gmail.com

Facebook: tatyana.bunto

Skype: tbunto89

Telegram: @Tbuntik