Что проверять при миграции данных, чтобы вовремя закончить проект

Эти проверки HFLabs использует на проектах системной интеграции. Чек-лист помогает найти ошибки данных в исходных системах, выгрузках из шины или ETL. Подходят для реляционных баз, во всю мощь раскрываются на объемах от миллиона клиентов.

Для тестировщиков, внедренцев enterprise-продуктов, системных интеграторов-аналитиков

H F Labs

+ 7 495 928-86-41
ask@hflabs.ru
www.hflabs.ru

119034, Россия, Москва, Турчанинов переулок, д. 6 стр. 2 офис 104, БЦ «Крымский мост»

Заполненность полей и null-значения

Сколько всего заполнено строк в таблице

```
SELECT count(*) FROM <table_name>;
```

Клиенты — физические лица	Количество
Всего	99 966 324

Проверять — Адекватно ли ожиданиям количество записей.

Сколько строк заполнено по каждому полю отдельно

```
SELECT <column_name>, COUNT(*) AS <column_name> cnt
FROM <table_name> WHERE <column_name> IS NOT NULL;
```

Физические лица	Количество
Всего	99 966 324
ДР	0
ИНН	65 136

Проверять

- Количество заполненных строк в каждом поле.
- Соотношение количества заполненных строк в каждом поле к количеству строк в таблице. Если оно слишком мало, это повод подумать, нужно ли тащить поле в целевую базу.

Изменение заполненности после каждой выгрузки

Физические лица	Выгрузка 1	Выгрузка 2	Дельта
Всего	99 966 324	94 847 160	-5 119 164

Проверять — Где прибавилось или убавилось.
Например, если между выгрузками исчезло 5 млн записей, это стоит изучить.

Популярные значения

Каковы топ-100 популярных значений в строковых полях

```
SELECT *
FROM
        (SELECT <column_name>, COUNT(*) cnt
        FROM <table_name>
        GROUP BY <column_name> ORDER BY 2 DESC)
WHERE ROWNUM <= 100;
```

Имя	Количество
-	541 727
Тест	333 789
Александр	192 834

Дата рождения	Количество
-	9 314 770
01.01.1900	117 078
15.09.2015	53 702

Проверять — Нет ли мусора и дефолтных данных среди самых популярных значений.

Какова популярность всех значений в полях-справочниках и классификаторах

```
SELECT <column_name>, COUNT(*) cnt
FROM <table_name>
GROUP BY <column_name> ORDER BY 2 DESC;
```

Место рождения	Количество
Россия	292 585
Россия	158 163
РОССИЯ	70

Проверять
- Для строковых полей-справочников — разницу в значениях.
- Для классификаторов — хватает ли значений.

Длина значений в строковых полях

Слишком короткие значения

```
SELECT * FROM <table_name>
WHERE LENGTH(<column_name>) <= 3;
```

ФИО
-
qqq

Проверять — Не мусор ли в коротких строках. Часто так и есть.

Поля, заполненные впритык

```
SELECT * FROM <table_name>
WHERE LENGTH(<column_name>) = 65;
```

Адрес
119034, город Москва, переулок Турчанинов, дом 6, строение 2, пом

Проверять — Нет ли значений, поместившихся не полностью.

Как значения распределяются по длине

```
SELECT LENGTH(<column_name>), COUNT(<column_name>)
FROM <table_name> GROUP BY LENGTH(<column_name>);
```

Длина строки	Количество строк
124	130
125	1100
126	70

Проверять — Нет ли подозрительных перекосов в распределении.

Консистентность и кросс-сверки

Связаны ли данные, которым положено быть связанными

```
SELECT COUNT(*)
FROM
(
        (SELECT <ID1> FROM <table_name_1>)
        MINUS
        (SELECT <ID2> FROM <table_name_2>)
);
```

Проверять — Сколько в связанных таблицах несвязанных записей.

Нет ли дублирования первичных ключей в разных таблицах

Если не проверить, при миграции айдишники могут конфликтовать. Такое случается, когда исходно клиентов хранят в нескольких таблицах и в целевой системе объединяют.

Проверять — Нет ли риска конфликтов между первичными ключами в разных таблицах.

Еще пара важных проверок

Нет ли латинских символов там, где им не место

```
SELECT <column_name> FROM <table_name>
WHERE REGEXP_LIKE(<column_name>, '[A-Z]', 'i');
```

Проверять

Все ли в порядке, например, с кириллическими фамилиями. В них очень любит забиваться латинская С.

Не затесались ли посторонние символы в строковые поля, предназначенные для цифр

```
SELECT <column_name>
FROM <table_name>
WHERE REGEXP_LIKE(<column_name>, '[^0-9]');
```

Проверять

Нет ли чего лишнего в ИНН и номерах паспортов, например. Или в телефонах, но с исключением для плюса, скобки и дефиса. И не встречается ли буква «О» вместо нулей в строковых полях, где хранятся цифры.

Насколько данные адекватны

С данными просто нужна революционная бдительность. Всегда.

Проверять

- 50 000 телефонов у клиента «Софья Владимировна» — это нормально?
- ИНН как бы заполнены, но в столбце лежит «79853617764», «4956780966» и т. д. Что за телефоны, окуда? Где ИНН?
- Поле «Адрес одной строкой» не соответствует полям, в которых адрес хранится по частям. Почему адреса разные?

С базовой аналитикой на этом все, изучайте данные!

Больше статей о данных: **blog.hflabs.ru**

H F Labs

+ 7 495 928-86-41
ask@hflabs.ru
www.hflabs.ru

119034, Россия, Москва, Турчанинов переулок,
д. 6 стр. 2 офис 104, БЦ «Крымский мост»