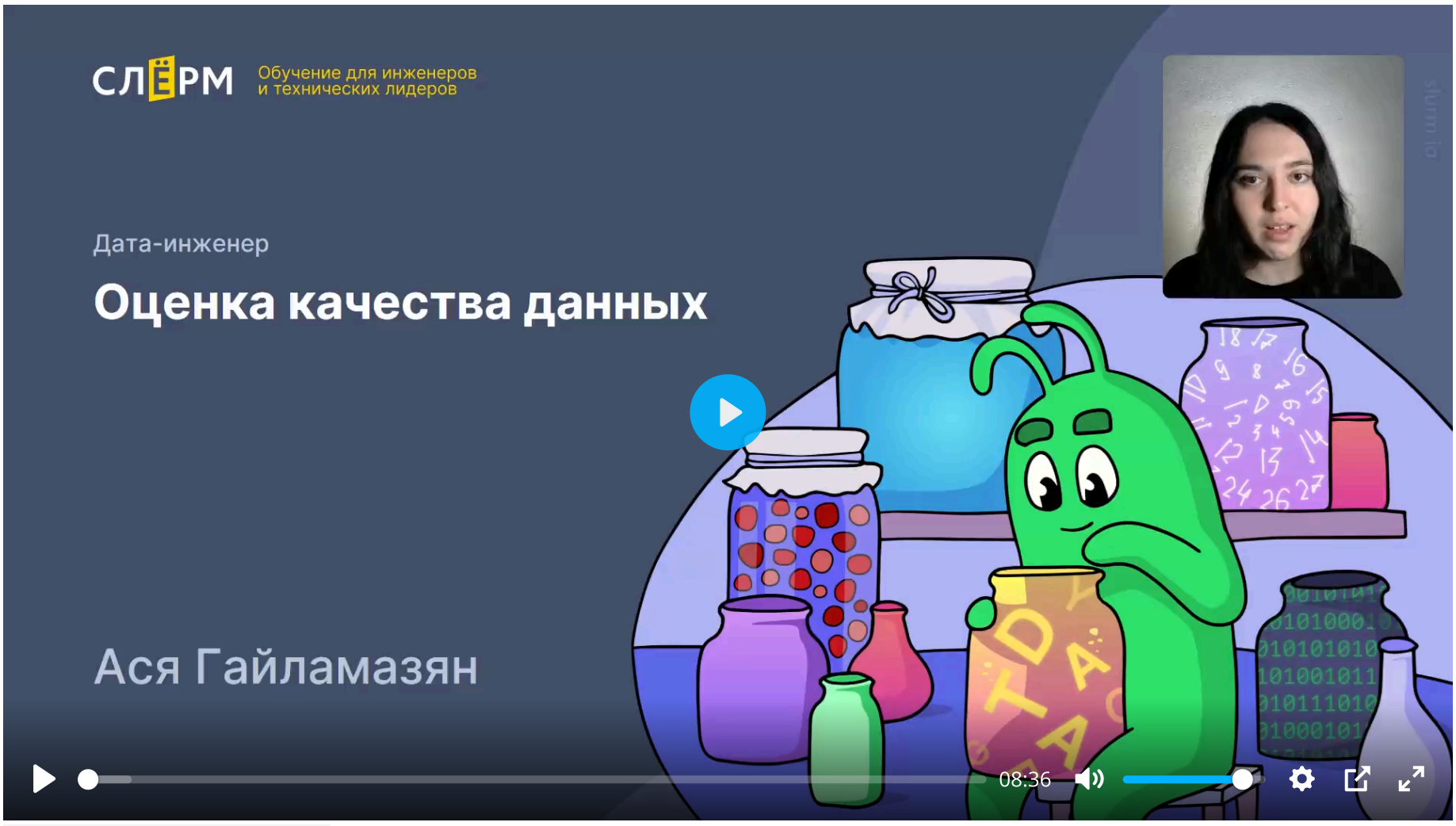


Презентация к уроку 12.2.4

Тексовая расшифровка видео:

ОЦЕНКА КАЧЕСТВА ДАННЫХ

План:

- Оценка качества данных;
- Точность, полнота, согласованность;
- Своевременность, актуальность, валидность;
- Целостность, дублирование, уникальность;
- Четкость, надежность, доступность;
- Возможность аудита, документация, этические соображения.

Оценка качества данных

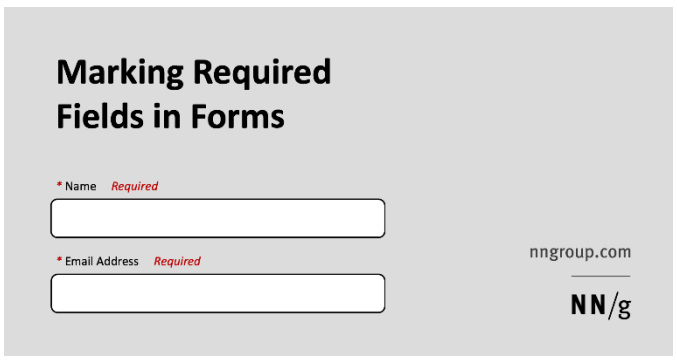
В компаниях есть определенные правила, принципы, технологии, которые объясняют, как обеспечиваются безопасность и качество данных. Поговорим о том, как оцениваются качества данных.

Точность, полнота, согласованность

Точность данных – критерий, сообщающий о том, что данные неошибочные, что они согласованы и не содержат несоответствий и расхождений.

Чтобы проверить точность данных, можно устроить периодические сверки данных на источнике и на данных из других мест. Так, например, телеком-операторы устраивают поверки коммутаторов (это оборудование, которое получает тарификационные и другие события) и сверяет их с теми данными, которые поступили из других источников, используя тестовый набор данных.

Полнота данных – критерий, при котором необходимые для анализа данные должны всегда поступать полностью. Так, например, при заполнении форм пользователями, можно помечать поля звездочками, отправлять пользователю email с подтверждениями.



Согласованность данных – критерий, сообщающий о том, что данные должны быть согласованными по форматам и т.д.

Table 2.1. Standard data input tools

Text input box	Pull down list	List	Radio box	Check box
Feb	Feb	Feb Mar Apr May Jun Jul Aug	1 2 3 4	1 2 3 4

Своевременность, актуальность, валидность

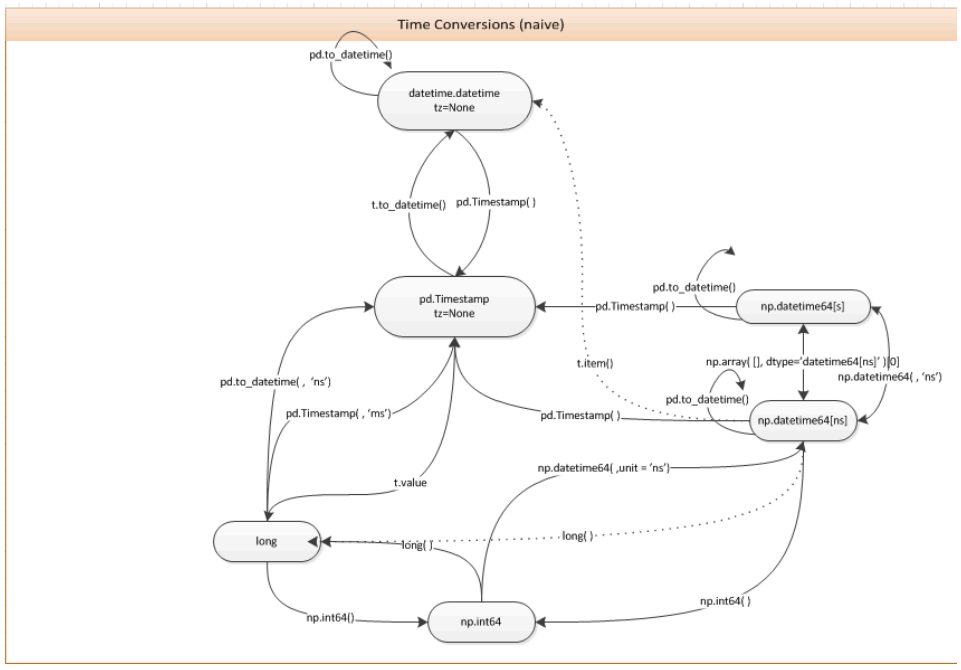
Своевременность данных – критерий, при котором данные получаются максимально быстро. Например, в Kafka есть оборудование, на котором можно смотреть:

- в какой момент на источнике появляются данные;
- в какой момент Kafka получает данные, когда те становятся доступными для обработки.

Если задержка слишком большая, то стоит задуматься о том, как это исправить. Для своевременности забора данных можно придумывать различные стратегии.

Актуальность данных – критерий, при котором берутся только необходимые для конкретного случая данные. Так, например, клиентов можно сегментизировать, после чего брать только тот сегмент, который необходим для

Валидность данных – критерий, при котором происходит проверка соответствия данных на определенный формат.



Целостность, дублирование, уникальность

Целостность данных – критерий, означающий, что к данным нет несанкционированного доступа.

Для этого можно:

- делать бэкапы;
- создавать резервные копии данных;
- обеспечивать Role-based access control, чтобы данные были доступны определенным пользователям;
- добавлять различные проверки действий над данными.

Дублирование данных – критерий, от которого следует воздерживаться, во избежание излишнего представления данных. Если данные будут представлены излишне, это приведет к неверным результатам их анализа.

Этого можно избежать с помощью сверки полей «в лоб» или других методов. Так, например, ClickHouse обеспечивает движки CollapsingMergeTree, ReplicatedMergeTree, которые схлопывают по повторяющимся определенным ключам строки.

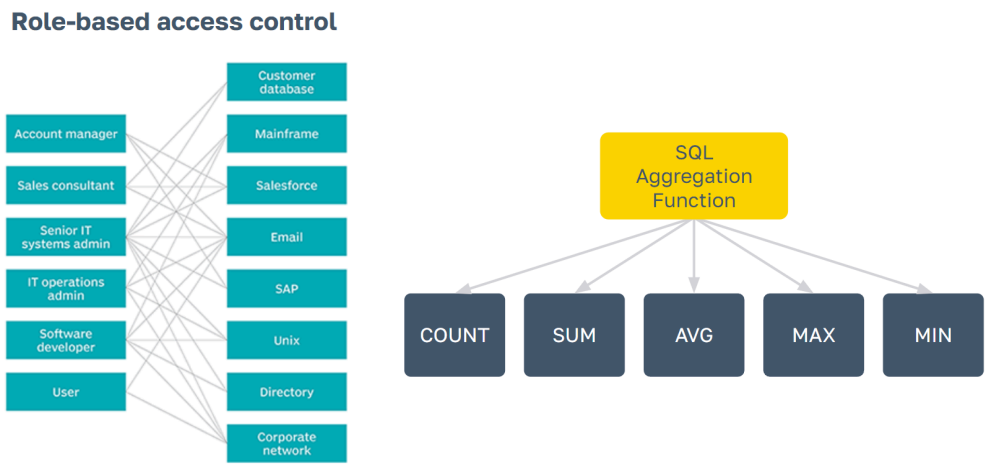
Уникальность данных – критерий, при котором каждый кусок данных должен быть обеспечен уникальным ID. Это можно делать на уровне БД, задавая поле как уникальное/неповторяющееся/ключевое, чтобы каждая строка имела свой идентификатор.

Четкость, надежность, доступность

Четкость данных – критерий, при котором каждому уровню анализа нужна определенная разбивка. То есть для некоторых видов анализа можно агрегировать данные и не представлять их излишне детализировано.

Надежность данных – критерий, при котором создаются резервные копии для получения данных. Например, в Kafka есть репликации, которые быстро восстанавливают получение данных с помощью надежных протоколов передачи данных и бэкапов.

Доступность данных – критерий, обеспечивающийся с помощью Role-based access control. Нужно выяснить кому, в каком объеме, с какими правами эти данные нужны. Для того, чтобы отслеживать правильные доступы к данным, можно завести логи доступа.



Возможность аудита, документация, этические соображения

Возможность аудита данных подразумевает под собой весь путь прохождения данных, который можно хранить в журнале логов (не только доступ пользователей/то, что ими производилось, но и преобразования, алгоритмы, математические формулы и т.д.).

Три простых шага аудита данных:

1. Привлеките заинтересованные стороны

Данные, важные для опыта клиентов, вероятно, хранятся и используются в разных отделах. Найдите и привлеките ключевые заинтересованные стороны, которые могут рассказать о процессах сбора, хранения и использования данных.

2. Составьте карту расположения ваших данных

Найдите все места, где хранятся данные, важные для обслуживания клиентов, и наметьте информационную архитектуру этих данных, включая то, как они хранятся и кто имеет доступ.

3. Оцените точность, широту и последовательность

Углубитесь и оцените качество своих данных, используя принципы точности, широты и последовательности, а затем проведите мозговой штурм по решению любых проблем, которые вы обнаружите в ходе аудита.

Этические соображения рассматриваются с различных сторон:

- **Со стороны государства.** Например, когда телеком-операторам поступает запрос от государства предоставить данные по какому-то абоненту. Если телеком-оператор не может предоставить эти данные, то ему могут отозвать лицензию. Это критично.
- **Со стороны клиентов.** Компании должны по Федеральному закону №152 информировать клиентов об использовании их данных, а также получать согласия и подробно описывать то, какие действия будут производиться над данными. Те компании, с которыми вы обмениваетесь данными клиентами, должны придерживаться тех же принципов этики.

Чтобы соблюдать этику данных можно их анонимизировать, то есть вместо конкретных логинов/email'ов/имен использовать ID; можно использовать надежные протоколы, чтобы данные никуда не утекали.

Данные проходят несколько слоев обработки: