

Построение системы контроля качества данных: от ТР до работы в реальной компании

Евгений Виноградов, ЮMoney

Предпосылки



Скрытое
недовольство



Открытое
недовольство



Неправильные
решения



Различные
кары

Способы решения

- Разрозненные действия/проверки
- Blockchain
- Регламент контроля качества данных
- ИС DQ
- Расширение покрытия



Эволюция требований

- ТКД
- Проверки
- Результаты проверок
- Data Steward



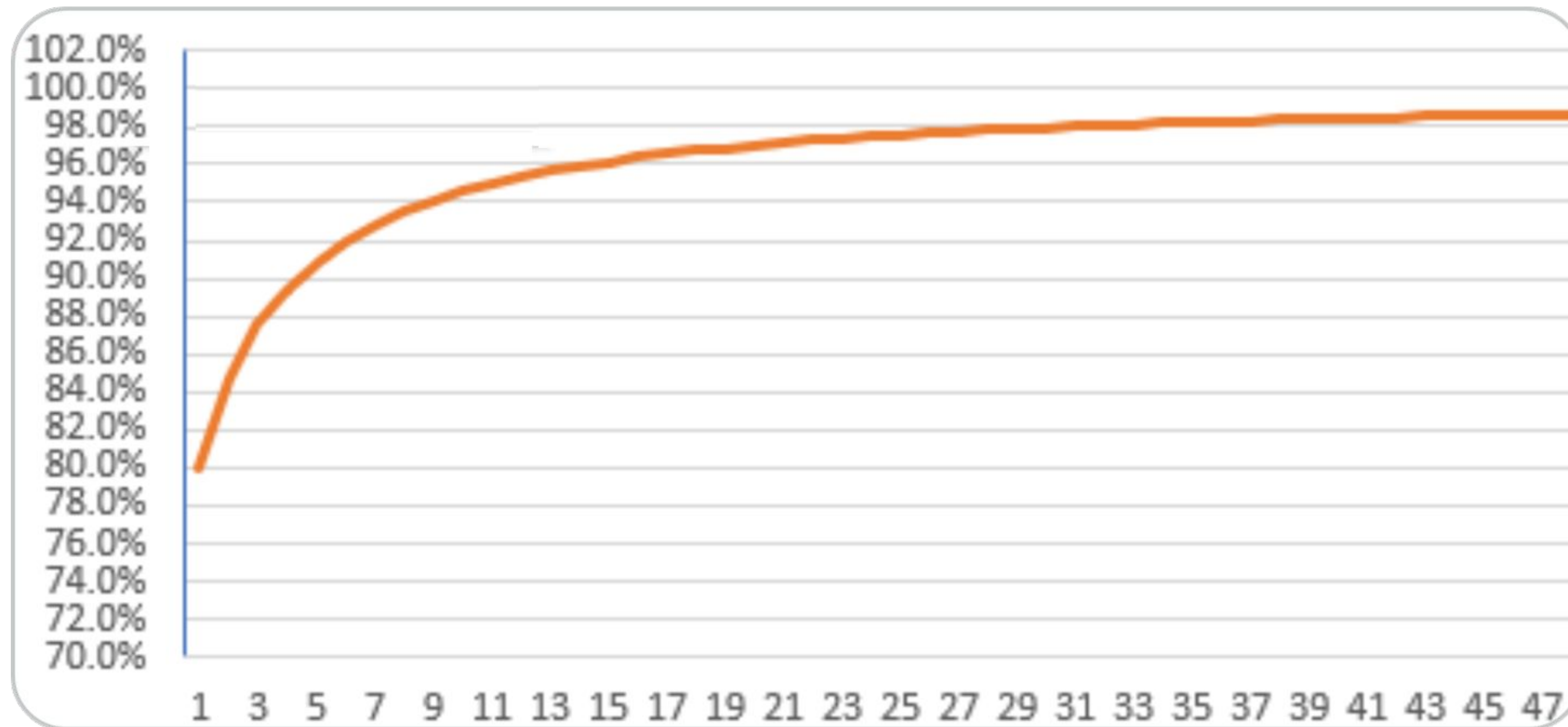
ТКД

Характеристики КД

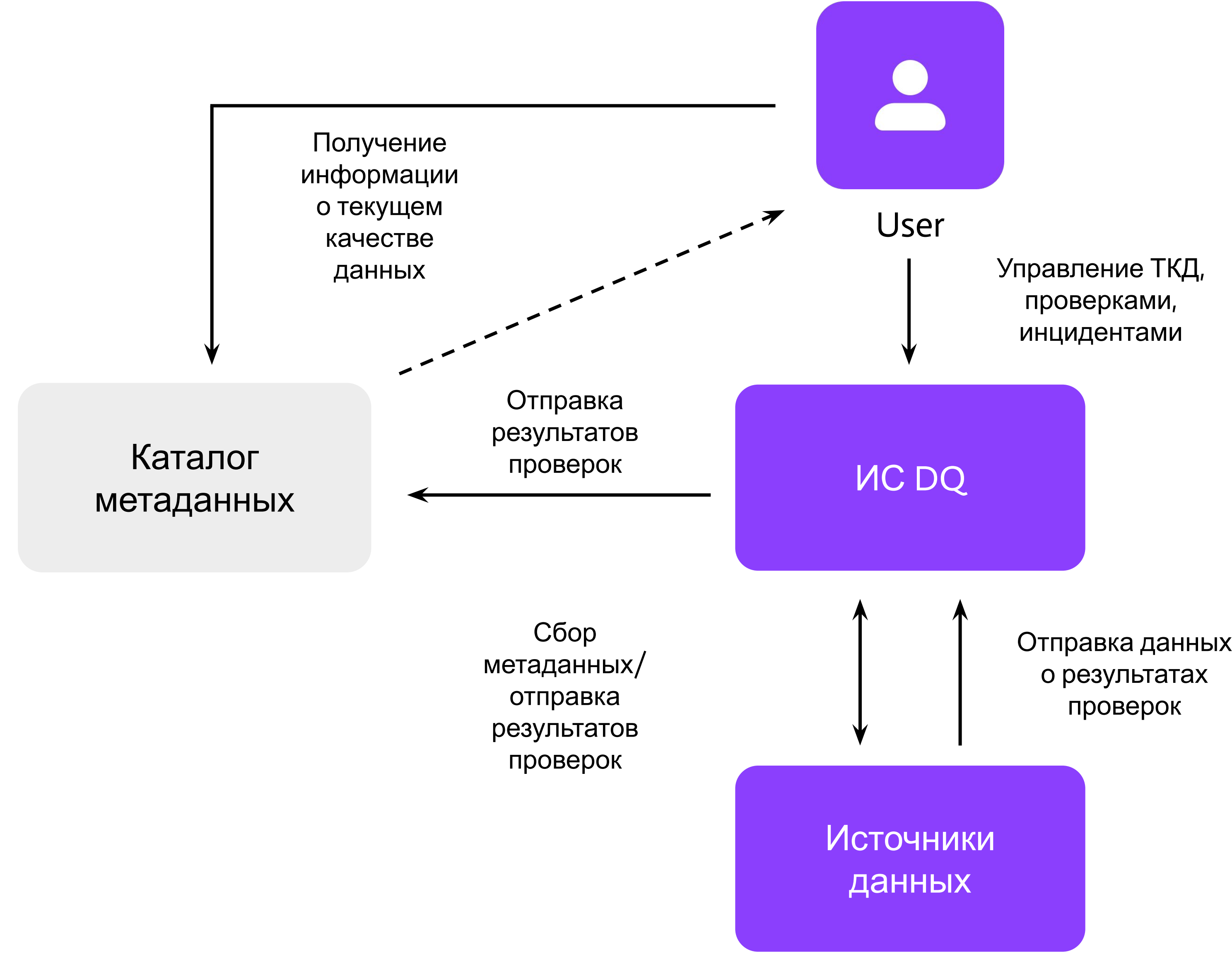
Характеристика	Описание
Достоверность	Соответствие реальным и статистически наиболее вероятным значениям свойств объектов реального мира
Точность	Отсутствие синтаксических и семантических ошибок
Полнота	Достаточность глубины (периода данных) и широты данных (охвата данными всех разрезов)
Актуальность	Уровень сохранения ценности данных в момент их использования
Согласованность	Взаимная непротиворечивость данных, хранящихся во внутренних АС, и согласованность с внешними источниками.
Уникальность	Уникальность данных в наборе выражается в отсутствии объектов-клонов или иных тождественных друг другу сущностей

ТКД

$$\text{Доля элементов без ошибок} = \frac{|R_{\text{correct}}|}{|R_{\text{total}}|}$$



Общая схема



Управление ТКД

DataQuality

Контроль качества данных

Hello, YAN

Реестр ТКД

Описание требований	Наименование продукта	Заказчик	Дата создания	Дата отключения
Мера "Количество плательщиков" не всегда проставляется, пустое значение	БГД	ВІ	23.08.2024 12:00:58	
В продуктах безопасная сделка - сплитование неверно считается ИТОГО (процент чистого дохода не должен превышать 2-3 процента, все что выше - аномалия)	БГД	В2В-Продукты	23.08.2024 12:00:58	

Где будет точка входа?

AMUNDSEN

<

Платежи ☆

Datasets • SSAS • EPR

Description

Add Description

Date Range

No valid partitions found
Unknown date range

Tags

DataQuality ssas
measure group epr

Owners

+ Add Owner

Frequent Users

No frequent users exist

[2] Проверка На Итого По Мпс

DQ State

Alert

Execution Date	Min	Target	Actual	State
09.12.2024 06:11:14	90.00	100.00	7.14	alert
08.12.2024 06:04:46	90.00	100.00	7.14	alert
07.12.2024 06:04:47	90.00	100.00	7.14	alert
06.12.2024 06:09:35	90.00	100.00	7.14	alert
05.12.2024 06:09:39	90.00	100.00	7.14	alert

Ключевые выводы

- Запуск системы контроля качества данных меняет образ мыслей бизнес-пользователя
- ТКД – ключевой элемент, но их не всегда могут сформулировать исчерпывающе
- Сама по себе DQ дает только замеры, а Data Steward и регламенты позволяют менять ситуацию



Спасибо!

jonny@yoomoney.ru