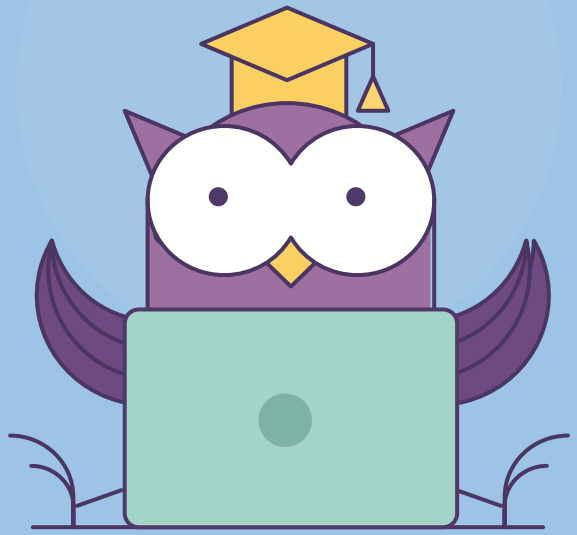
Data Quality. Контроль качества данных, мастер-данные.

Козырь Артемий Сергеевич

Не забыть включить запись

Меня хорошо слышно && видно?



Напишите в чат, если есть проблемы!

Ставьте + если все хорошо
Ставьте - если есть проблемы

Правила вебинара

Активно участвуем

Задаем вопрос в чат или голосом

slack

Off-topic обсуждаем в Slack #канал группы или #general

Вопросы вижу в чате, могу ответить не сразу

План занятия

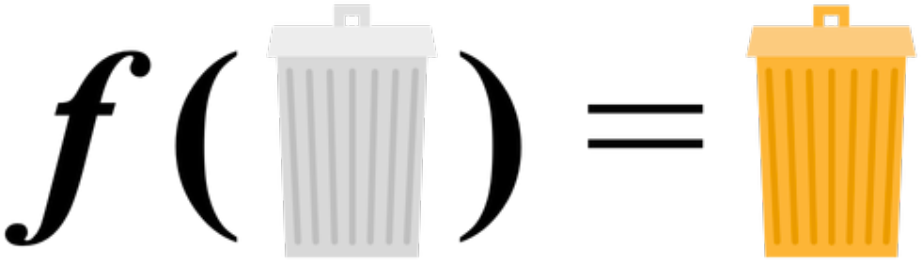
Качество данных и метрики	Причины, примеры и управление КД	Измерение, мониторинг, исправление	MDM

Качество данных

- Оценка пригодности данных
- Для заданных целей и задач

Зачем нужно управление! КД

- Garbage In – Garbage Out



На что влияет КД

- Доверие используемым данных
- Качество управленческих решений
- Эффективный маркетинг
- Удовлетворенность клиентов
- Снижение затрат / повышение маржинальности
- Регуляторные и репутационные риски

Оценка качества данных

- **Completeness**: пропуски в данных?
- **Validity**: соответствие правилам?
- **Uniqueness**: наличие дубликатов?
- **Consistency**: согласованность между наборами данных
- **Timeliness**: актуальность на момент времени
- **Accuracy:** данные отражены верно (соответствуют действительности)

Completeness

- Полнота
- Заполнены все атрибуты

Имя	Адрес	Телефон
Андрей	Нагатинская 10	79155007080
Артур		79158203040

Consistency

- Согласованность между наборами данных
- Разные пользователи получают одинаковую информацию (согласованную)
- Карточки с истекшим сроком действия не могут быть в статусе Active и иметь активность по счету

Имя	Адрес	Телефон
Андрей	Нагатинская 10	79155007080
Артур		79158203040

Referential integrity

- Ссылочная целостность
- Справочники и связи

Дата	Клиент	Продукт
2019-01-01	1	1
2019-01-01	1	2

Продукт	Наименование	Цена
1	Арахис	100

Domain Integrity

- Пол: М / Ж
- Возраст
- Страна проживания
- Всевозможные коды: ОКВЭД, ...

Бизнес-правила

- Баланс: дебет и кредит
- Обороты по счетам
- Один департамент – один начальник
- Один проект – один менеджер

Timeliness - актуальность

- Посылка доставлена 1 января, но данные в хранилище обновились 2 января
- Каким числом данные отразятся в отчете?

Дата	Количество посылок
2019-01-01	0
2019-01-02	1

Проблемы качества данных

Код	Наименование	Страна	Валюта	Дата погашения	Рейтинг	Вероятность дефолта (PD)
34598	Bank of Scotland	GB	GBP	01.12.2012	AA**	0,3
65656	Химмашимпэкс	RUS	RUB	21.02.2010	B	4,0
54335	ООО "Издательство "ВОКРУГ СВЕТА"	RUS	RUB	17.06.2011	AA	0,3
45667	Волгоэлектромонтаж	RUS	RUB	18.01.2010	AA	0,3
32345	ЗАО "Воскресенские тепловые сети"	RUS	RUB	19.01.2010	2A	0,3
8468	ООО"Технострой" (филиал в Воронеже)	RUS	RUB	18.03.2010	B	4,0
75435	ГУП "Мосзеленхоз"	RUS	RUB	02.01.2011	AA	0,3
	ООО "Навигатор"	RUS	RUB	15.01.2011	CC	0,3
23411	Delaware Bay Company	US	USD	23.01.2007	B	4,0
64721	First Allied Securities	US	EUR	24.02.2010	B	104,0
98723	ЗАО "Первоуральский торговый дом"	RUS	RUB	26.08.2010	B	4,0
ASD43	ООО "Печатный Мир"	RUS	RUB	27.01.2010	B	4,0
83256	ОАО "Плутон"	RUS	RUB	28.01.2010		4,0
5185	ОАО "Полимербыт"	RUS	RUB	29.01.2010	B	4,0
51623	#INPUT ERROR	RUS	RUB	29.12.2010	B	-4,0
45	ООО "Провинция-2000"	RUS	RUB	30.01.2010	B	4,0
278	ООО"Технострой" (филиал в Москве)	RUS	RUB	14.01.2010	B	4,0
167	ЗАО "Топливно-бункерная компания"	RUS	RUB	01.02.2010	B	4,0
2086	ТОВ компанія УКРТОРГСЕРВІС	UKR	RUB	02.02.2010	B	4,0
52457	НПО Укроргсинтез	UKR	UAH	05.05.2010	1B	4,0
43344	ООО "Фудстар"	RUS	RUB	05.02.2010	B	4,0
23323	ООО "Химкомплект" (НЕ ИСПОЛЬЗОВАТЬ!)	RUS	RUB	27.07.2010	B	4,0
54545	ОАО "Химконверс"	RUS		07.02.2014	B	4,0
87434	ЗАО "Химмашимпэкс"	RUS	RUB	21.02.2010	B	4,0
76788	ОАО "Хлебный Дом"	RUS	RUB	12.02.2010	B	4,0

- Полнота
- Соответствие стандартам
- Взаимное соответствие
- Дублирование
- Связность и целостность
- Корректность

План занятия

Качество данных и метрики	Причины, примеры и управление КД	Измерение, мониторинг, исправление	MDM

Примеры проблем с КД

- Розничный бизнес: множество телефонных номеров содержащих 000000000 или номера авиабилетов
- Компания в сфере здравоохранения: 9 различных значений в атрибуте «Пол»
- Компании проката авто: дубликаты договоров
- Могут стоить компании миллионы $$
- GDPR и 152-ФЗ (Персональные данные)

Некачественные данные

- Пропуски (NULL)
- Ошибочные данные
  - Неправильные типы данных, дубли
  - Нарушения ACID (Dirty Read, Non-Repeatable, Lost Update, Loss of transaction)
  - Ошибки в датах и категориальных данных
- Непригодные данные
  - Противоречивые (в разных источниках)
  - Двусмысленные
  - Конкатенация, специальные символы, порядок слов, аббревиатуры, и т.д.

Причины возникновения

- Человеческий фактор
- Ошибки в коде (баги)
  - На источнике
  - На стороне хранилища
- Инциденты: падения, восстановление, потеря транзакций, незавершенные вычисления
- Изменения в системах-источниках

Еще причины возникновения

- Бизнес-правила могут конфликтовать (коллизии)
- Невозможно получить все изменения в источнике (Change Data Capture)
- Ошибки в обработке SCD
- Ошибки в построении ETL-процессов
- Невозможность восстановить (рестартовать) ETL-процесс с чекпоинта без потери данных

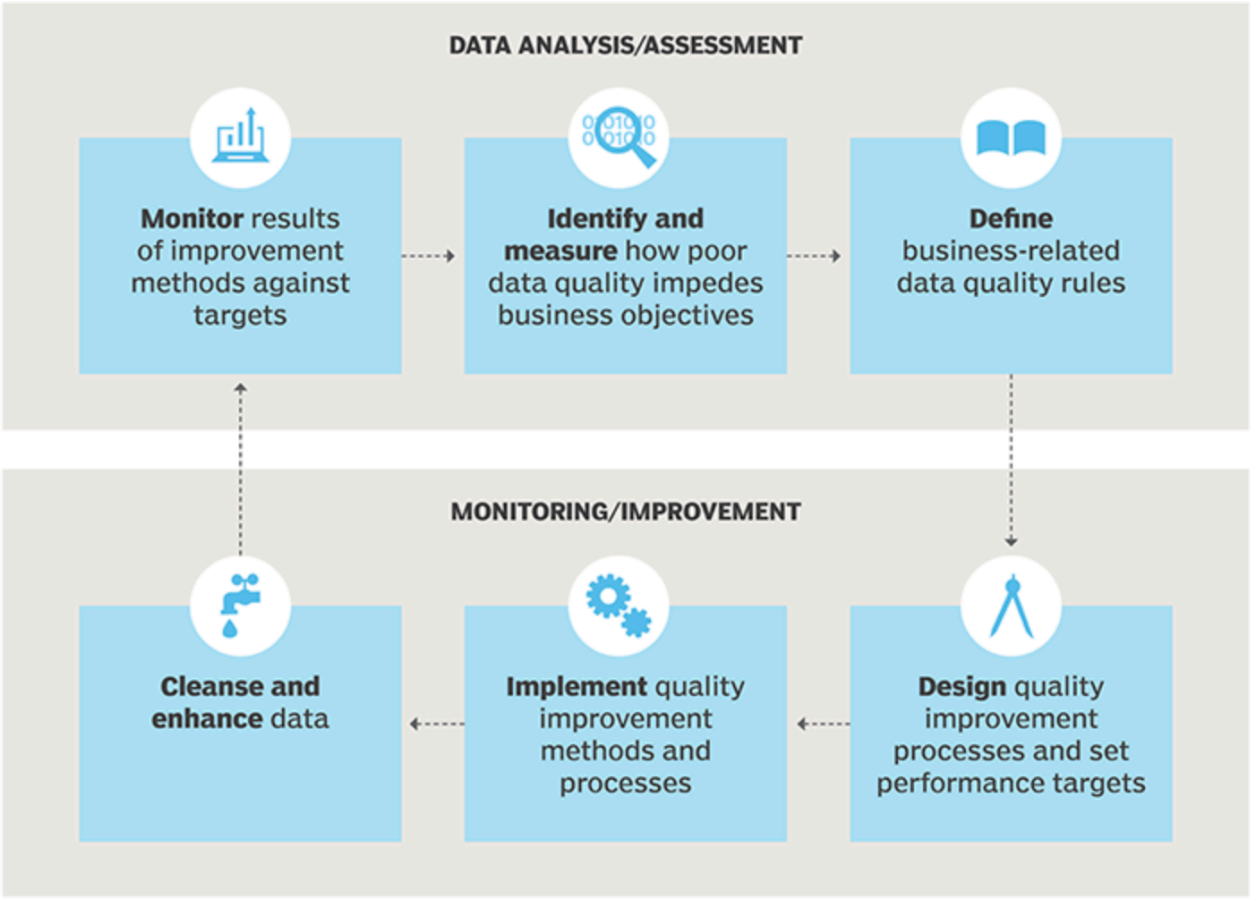
Как обеспечить КД

- **Организационный уровень**: роли, ответственность, стандарты, регламенты
- **Процессный уровень**: лучише практики, циклы Plan-Do-Check-Act
- **Технологический уровень**: программная реализация и архитектура решений

Роли в управлении КД

- **Data owner**: Владелец данных (ИТ-системы)
- **Data steward**: правила, требования к поставке данных
- **Data manager**: технологическая архитектура
- **Data users**: конечные пользователи

Цикл управления качеством данных



План занятия

Качество данных и метрики	Причины, примеры и управление КД	Измерение, мониторинг, исправление	MDM

Как измерить КД

- Data profiling (Профилирование данных)
- Изучить данные, найти проблемы, сформулировать выводы
- Систематическое измерение на заданные метрики
- Сопроводить процессы загрузки и расчетов измерениями метрик качества данных

Техническое и бизнес-качество

- **Техническое качество данных:**
  - Пропуски в данных, NULL, default values
  - Написание, грамматические ошибки, опечатки
  - Данные в неверном формате (ошибки в колонках)
  - Аномалии, выбросы
- **Бизнес-качество данных:**
  - Отчеты и витрины отвечают правилам бизнеса
  - Сальдо, баланс, дебет и кредит
  - Поставки и отгрузки производятся в срок
  - Все клиенты получили бонусы и скидочные предложения

Примеры проверок

- Соответствие схемы источник-приемник
  - Типы данных, длина, форматы, названия колонок
- Наличие пропусков (NULL / NOT NULL)
- Ограничения (constraints): PK, FK, Constraints
- Количество записей источник – приемник
  - Количество уникальных ключей
  - Rejected rows
  - Дубликаты
- Кросс-валидация, бизнес-правила

Как реагировать (обрабатывать)

- Отвергнуть строки с ошибками (reject rows)
- Попытаться сконвертировать
- Обрезать данные (trim)
- Удалить дубликаты
- Заменить на NULL для несоответствующих типов (non-numeric – numeric)
- Обрабатывать ограничения целостности (не грузить, присваивать «-1»)

Как исправить

- Data fix
- Выполнить исправительную загрузку
- Заполнить пропуски и дыры
- Удалить дубликаты (слить – merge)
- Восстановить историчность атрибутов
- **Предотвратить легче чем исправить**

Мониторинг КД

- Систематическое измерение на заданные метрики
- Сопроводить процессы загрузки и расчетов измерениями метрик качества данных
- Отчетность по качеству данных
- Исправление выявленных проблем

Мониторинг и нотификация

- Регулярный запуск оценки качества данных (по расписанию)
- Запуск по триггеру (после завершения расчета витрины)
- Ручной запуск
- Выслать результаты заинтересованным сторонам

Мониторинг и нотификация

- Записывать метрики качества данных в специальную таблицу
- Положить ошибочные строки в отдельную таблицу
- Ошибка, идентификатор записи, дата обнаружения, дата исправления

Что еще может КД

- Подсветить особые кейсы,
- Привлечь внимание к определенным проблемам
- Послужить началом изменений (процессов)

План занятия

Качество данных и метрики	Причины, примеры и управление КД	Измерение, мониторинг, исправление	MDM

Мастер-данные (MDM)

- Зачем нужны
- Какие бывают
- Клиентский MDM (CDI) и Data Quality

Какие бывают MDM-системы

- RDM (Reference Data Model) – фиксированные справочники. Редко меняются. Оргструктура.
- PDM (Product Data Model) – продукты, товары, материалы, услуги.
- CDI (Customer Data Integration) – клиенты, контрагенты.

Зачем нужен CDI?

- Сколько у нас клиентов?
- Кто из них приносит бОльшую часть прибыли?
- А кто убытки?
- Какие продукты востребованы?
- Лояльность
- Интересы клиента

Клиентский МДМ в банках

- Много систем с клиентскими данными
- Единый профиль клиента
- Слияния банков
- Физики и юрики
- Регуляторная и надзорная политика

Клиентский МДМ в телекоме

- От сим к клиенту
- Сведение биллинга
- Различные продукты и системы

Клиентский МДМ в страховых

- От полиса к клиенту
- Обогащение данных (уже заполненных в полисе)
- Агентские и партнерские взаимоотношения

Входные данные

Сергей В. Иванов
31.07.1983
Дубна 2-й Театральный
8(916)151-07-23
3-56-03

ИВАНОВ СЕРГЕЙ
1-I-2016
096213-56-03
ivanov_s@mycorp.ru
3-56-03

sregey vlodimirovich
ivanov
Дубна, Татральный 2, 6 17
0079161510723-моб
4957371293доб101

Целевая запись

Сергей Владимирович Иванов
31.07.1983
141983, Московская обл. г. Дубна проезд Театральный 2-й д. 6, кв. 17
+7 (49621) 3-56-03 +7 (916) 151-07-23 +7 (495) 737-12-93*101
ivanov_s@mycorp.ru

Разметка данных

- Загрузить (интегрировать)
- Стандартизировать
- Найти дубли
- Актуализировать (эталон)
- Удалить дубли

Методы выявления дубликатов

- Строгая логика: значения совпадают
- Вероятностная логика: допустимы ошибки в нескольких символах, перестановка букв, слогов
- Нечеткая логика

* Проблемы матчинга и как можно с ними бороться

Стратегии эталонизации

- Выбор системы-эталона
- Формирование золотой записи (из разных источников)

Стратегии эталонизации

- Выбор системы-эталона
- Формирование золотой записи (из разных источников)

	Карточка АБС ID=10	Карточка CRM ID=20
ФИО	Травин Иван	Травин Иван Сергеевич
Телефон	(495) 960-42-42	960-42-42
Дата рождения	31.03.1984	
Документ	7701 359743	7701 359743

Выбор мастера

	Карточка АБС ID=10	~~**Карточка CRM ID=20**~~
ФИО	Травин Иван Сергеевич	~~Травин Иван Сергеевич~~
Телефон	(495) 960-42-42	~~960-42-42~~
Дата рождения	31.03.1984	-
Документ	7701 359743	~~7701 359743~~

Золотая запись

	Карточка АБС ID=10	Карточка CRM ID=20	Золотая (хид + идентификаторы)
ФИО	Травин Иван	Травин Иван Сергеевич	Травин Иван Сергеевич
Телефон	(495) 960-42-42	960-42-42	(495) 960-42-42
Дата рождения	31.03.1984		31.03.1984
Документ	7701 359743	7701 359743	7701 359743

Демо

1. Удаление дублей на ПРОМ стенде хранилища
2. Восстановление истории в таблице после неудачного патча (баг)
3. Добавить проверки качества данных в небольшой pipeline
    1. Rowcount(source) == rowcount(target)
    2. Файл зарегистрирован
    3. Количество записей-дубликатов
    4. Количество записей, поступивших в детальный слой

Ссылки на материалы

- Работа с мастер-данными. Очистка клиентских данных (RT Data talks)
- Почему в Siebel CRM не получится вести единую базу клиентов
- Работа с качеством данных. Профилирование, очистка и DQ мониторинг.
- Проблемы матчинга и как можно с ними бороться
- DataQuality for BigData (github)
- Spark package for checking data quality (github)

Рефлексия

- Что вам запомнилось больше всего
- Пройти опрос

Ваши вопросы?