

Modern Storages and Data Warehousing Week 9 - Data Governance

Попов Илья, i.popov@hse.ru

1 - Homework Q&A

2 - Homework #3

Ресар прошлых занятий

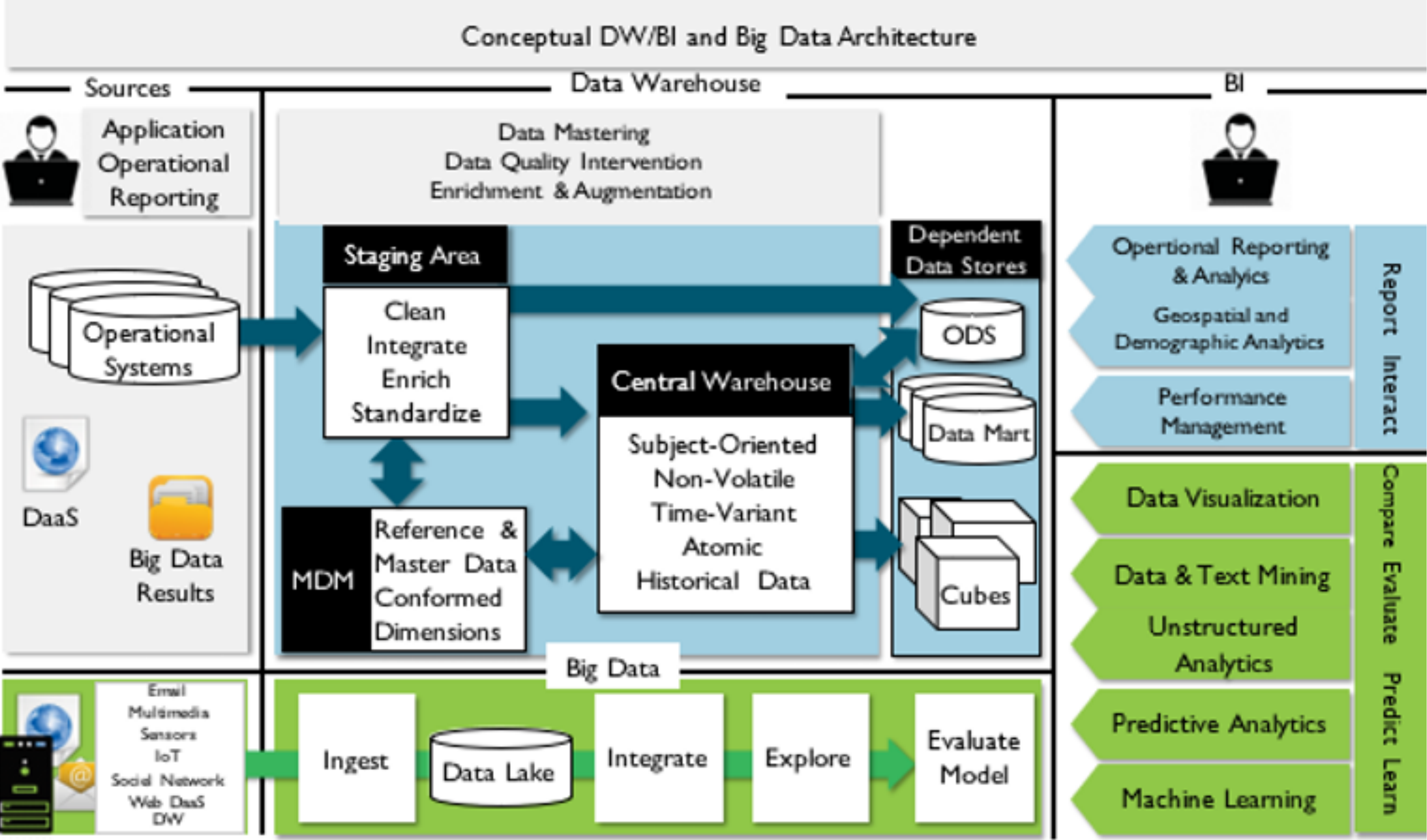
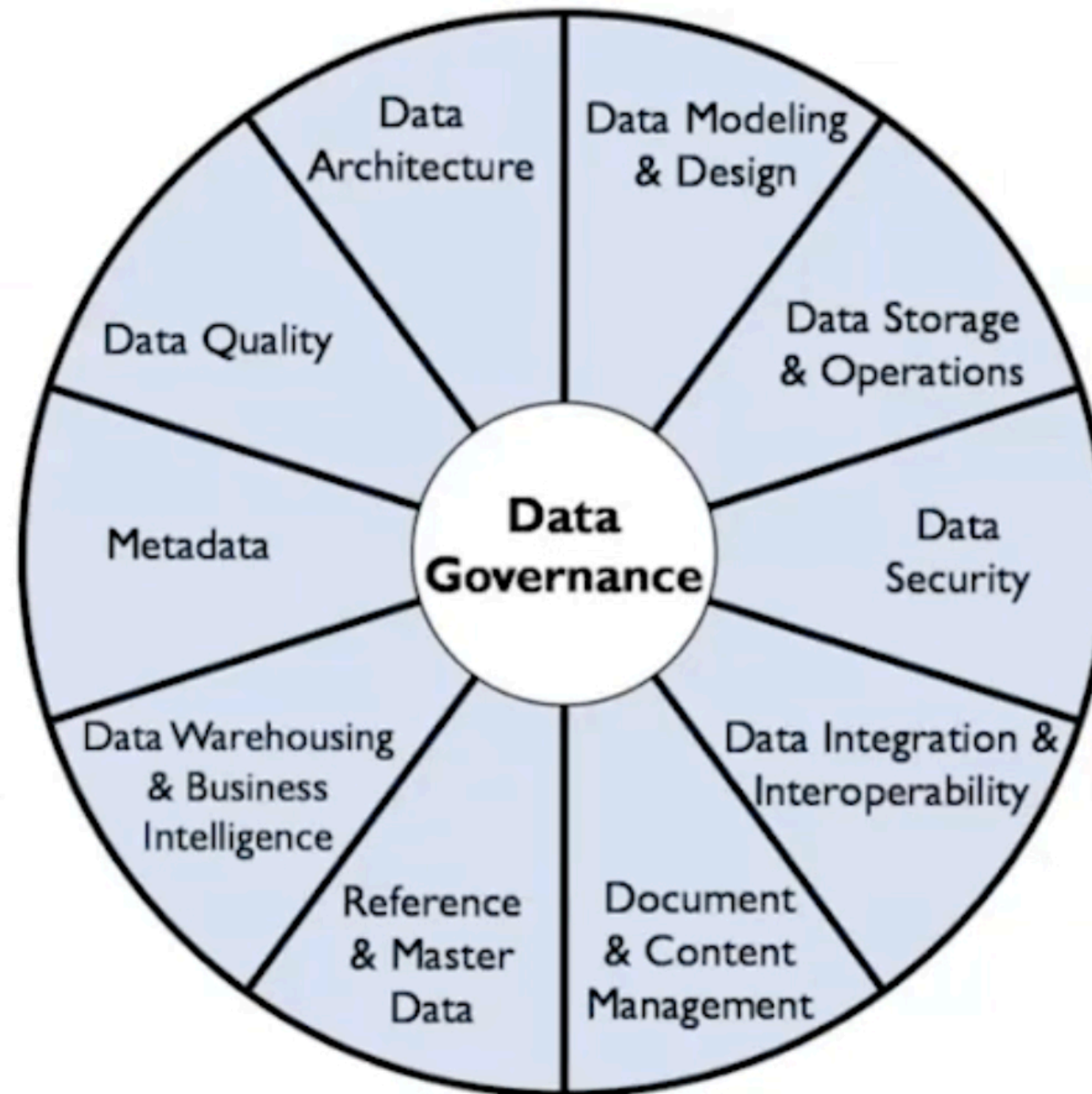



Figure 5: Data Warehouse Concept

Мотивация

- › Вот мы с вами построили (надеюсь) хранилище
- › Теперь это хранилище должно жить
- › Его нужно поддерживать, обновлять, управлять доступами к данным
- › Это все называется общим словом - Data Governance

Data Governance



 Доверие к данным -
краеугольный камень любого
DWH

Ральф Кимбалл

Data Quality

Качество данных — это степень, с которой набор характеристик, присущих данным, отвечает конкретным требованиям с точки зрения их применения.

Неправильно выстроенные уровни качества данных непосредственно влияют на успех проекта: можно либо задать слишком высокий уровень и не достигнуть его никогда, либо установить слишком низкий и будет потерян смысл системы аналитики.

Управление качеством данных — согласованная деятельность по контролю и управлению структурой, имеющей непосредственное отношение к качеству данных, обеспечение соответствия данных целям их использования с поддержанием полноты, точности, корректности и своевременности.

Критерии качества данных

Единых критериев для качества данных не существует

Критерии качества данных по DAMA (DAta Management Association):

- › **Полнота** - отношение фактически имеющегося в хранилище объема данных к потенциально доступному
- › **Уникальность** - ни одному реально существующему объекту не должно соответствовать более одной записи
- › **Актуальность** - степень отражения реального положения дел в момент времени
- › **Годность** - синтаксическое соответствие данных
- › **Соответствие** - степень соответствия данных реальным объектам
- › **Согласованность** - отсутствие противоречий в данных

Критерии качества данных

Что еще можно сюда добавить?

- › **Полезность** - насколько понятны данные
- › **Своевременность реагирования** - насколько оперативно мы можем внести изменения в данные
- › **Гибкость** - насколько данные сопоставимы с другими данными
- › **Надежность** - какова репутация / цена ошибки в данных
- › **Ценность** - есть ли экономическая целесообразность / окупаемость затрат на владение данными

Критерии качества данных

При этом, в DAMA DMBOK критерии другие:

- › **Актуальность** - сроки получения (синхронизация / отставание / волатильность)
- › **Консистентность и допустимость** - соответствие заданным значениям
- › **Полнота** - соответствие данным в источнике всем ожидаемым
- › **Разумность** - “соответствие здравому смыслу”
- › **Согласованность** - нет противоречий внутри себя
- › **Соответствие** - близость данных к реальности
- › **Ценность** - экономическая целесообразность
- › **Уникальность** - отсутствие дублирования данных внутри слоя

Целостность данных

- › В широком смысле - полнота, точность, согласованность
- › В узком смысле - целостность данных на уровне ссылок
- › Сироты - ссылки, ведущие в никуда
- › Дубли - полностью идентичные строки

Проблематика

- | Ответа на вопрос, что такое качественные данные - нет
- › Степень качества данных определяется по тому, пригодны или непригодны они к использованию
- › Использование всегда рассматривается в контексте потребителей данных
- › Потребители не могут сформулировать потребность в качестве данных

Оценка качества данных

- › **Единовременная оценка качества** - производится для некоторого массива данных один раз, после чего в зависимости от результатов к данным применяются те или иные методы очистки
- › **Мониторинг** - потоки данных, поступающие в хранилище и далее в аналитику, непрерывно сканируются в поисках ошибок и несоответствий
- › **Визуальная оценка** - делается в аналитических / BI системах

Причины некачественных данных

- › Проблемы вследствие недостатка лидерства
- › Проблемы в результате ввода данных
- › Проблемы на стадии обработки данных
- › Проблемы, обусловленные системными / проектными решениями
- › Проблемы вследствие непродуманного исправления предыдущих проблем

Проблемы с лидерством

Недопонимание на уровне руководства компании ценности данных

- › Очень мало организаций управляют своими данными как ценным активом, еще меньше делают это с должной тщательностью
- › Рассогласованность данных внутри более серьезная проблема, чем ошибки в данных
- › Единообразие технологии внутри компании упрощает интеграцию данных в хранилище
- › Недопонимание руководителями ценности управления данными как активом приводит к снижению активности в сфере обеспечения их качества

Проблемы в результате ввода данных

- › Плохо запроектированный режим ввода
- › Слишком длинные / неупорядоченные списки
- › Переназначение полей. Использование старых полей для новых значений
- › Человеческий фактор
- › Изменение в бизнес-процессах не отражено в системах
- › Рассогласованность в бизнес-процессах

Проблемы на стадии обработки данных

- › Неверное представление об источнике
- › Устаревшие бизнес-правила
- › Изменение в структуре данных

Повторные проблемы при исправлении проблем

- | Ручная правка данных на проде — ЗЛО
- › Ручные скрипты с изменением данных часто используются для очистки данных
- › Неправильный код ведет только к большему количеству ошибок
- › Такие изменения сложно откатить

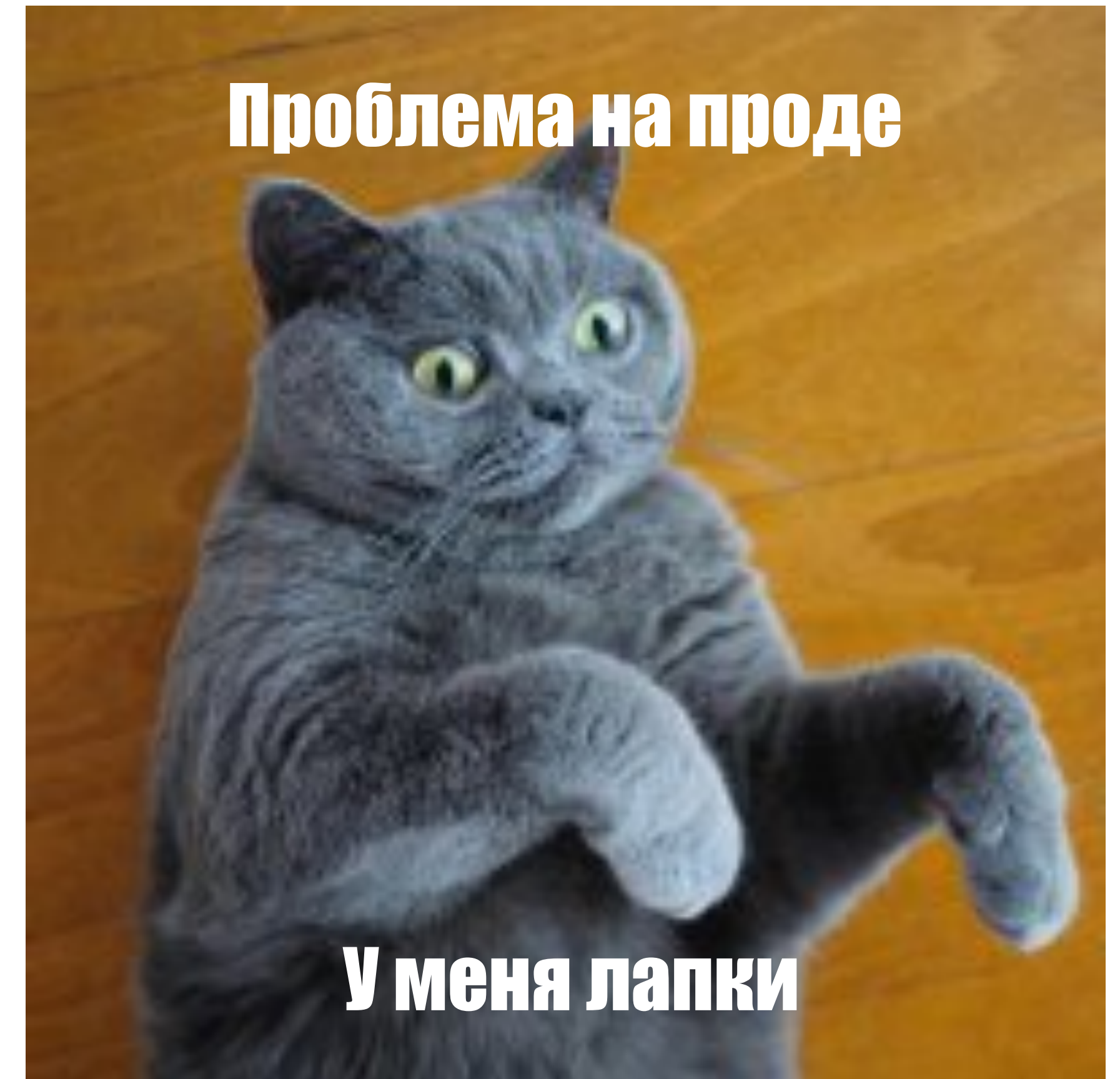
Где производить очистку данных

› В умах руководства



Где производить очистку данных

- › В умах руководства
- › Непосредственно в источниках данных



Где производить очистку данных

- › В умах руководства
- › Непосредственно в источниках данных
- › В ETL-процесса
- › Непосредственно в аналитическом контуре

Уровни качества данных

- › В умах руководства
- › Непосредственно в источниках данных
- › В ETL-процесса
- › Непосредственно в аналитическом контуре

Уровни качества данных

- › **Технический уровень** - на качество данных влияют в основном факторы, связанные с нарушением структуры, целостности, полноты, некорректные форматы данных.
- › **Аналитический уровень** - факторы, мешающие выполнить корректный анализ данных и получить достоверные результаты. Шумы в данных, аномальные значения, противоречия, дублирование записей и т.д.
- › **Концептуальный уровень** - проблемы с тем, что была выбрана неверная стратегия сбора данных. Собранные данные содержат недостаточно информации для описания БП.

Уровни качества данных

Уровень	Факторы	Проявление	Место борьбы
Технический	<ul style="list-style-type: none">Нарушение в структуре данныхНекорректное именование таблиц и полейНекорректные форматы кодирования данныхНарушение полноты и целостности данных<u>Противоречия</u> и дубликаты	Мешают интегрированию и загрузки данных в аналитические системы	Источники данных, ETL
Аналитический	<ul style="list-style-type: none">ПропускиАномальные значенияШумыПротиворечия и дубликаты на уровне записей	Снижают достоверность данных и искажают результаты	Источники данных, ETL, аналитические системы
Концептуальный	<ul style="list-style-type: none">Собранные данные не отражают бизнес в полной мере	Недостаток данных для анализа	Стадия проектирования хранилища

Профилирование данных

В рамках профайлинга данных проверяется соответствие атрибута наложенным на него ограничениям. Выполняется на основе метаданных, описывающих структуру данных.

- › **Тип атрибута.** Если тип атрибута определен, как числовой, а при проверке обнаружено, что он имеет другой тип, то выясняются причины и производитель соответствующее преобразование данных.
- › **Длина значения.** Определяется максимально допустимое количество символов в значении поля.
- › **Дискретные значения.** Проверяется частота и уникальность.
- › **Диапазон допустимых значений.** Задаются минимальное и максимальные значения, которые может принимать атрибут
- › **Анализ строковых шаблонов.**

Очистка данных

Очистка данных заключается в их преобразовании с целью приведения в соответствие с требованиями стандартов

Очистка данных в ETL процессе – дорогостоящее занятие. Для снижения затрат на очистку данных необходимо:

- › Внедрение защиты от некорректного ввода в системах источниках
- › Исправление данных в системах источниках
- › Совершенствование бизнес-процессов

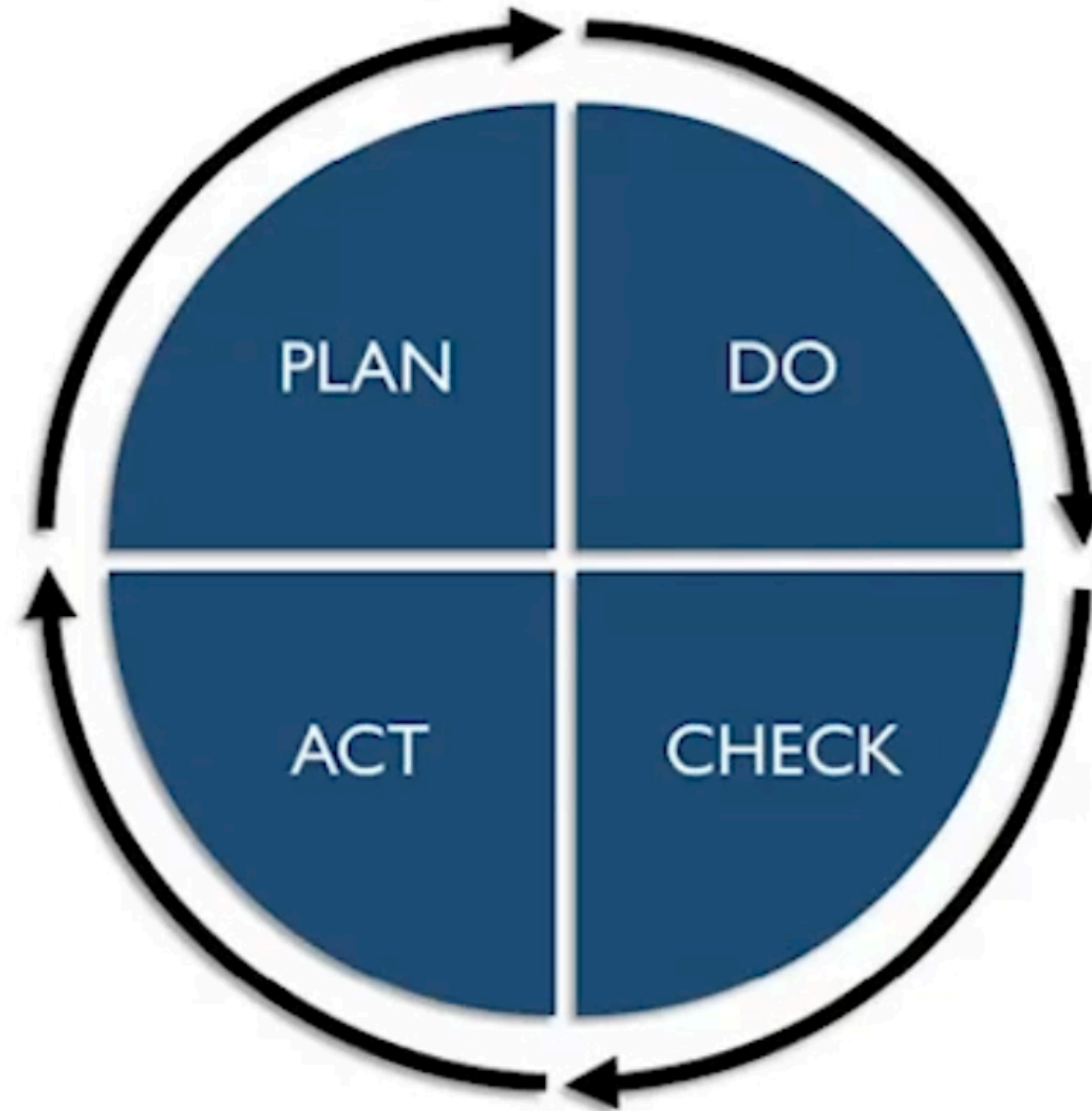
Улучшение качества данных

- Улучшение (обогащение) данных состоит в добавлении к набору атрибутов, повышающих качество данных
 - › Метки даты\времени: технические поля загрузки данных в ХД
 - › Аудит данных: добавление меток системы источника
 - › Справочные словари: приведение данных к стандартной для бизнеса терминологии

Мониторинг качества данных

- | Средства мониторинга качества данных зависят от самого определения качества данных в компании
- › Соответствие метаданным. Для каждого источника мы знаем структуру данных и можем оценить, получили ли мы то, что ожидали
- › Профайлинг данных
- › SLA доезда / недоезда данных
- › Соответствие данных друг другу / источнику
- › Статистический профайлинг данных

Процесс наведения порядка



Процесс наведения порядка

- › Определение качества данных
 - › Определение стратегии качества данных
 - › Определение критически важных данных для БП
-
- › Проведение первичной оценки данных
 - › Формирование целей по улучшению данных
 - › Приоритезация конкретных улучшений

Где все это ставить?

- › Между источником и raw - сверки с источником
- › Между raw и stage - профайлинг и исправление ошибок
- › При сборке stage - повышение качества данных
- › При сборке DDS - консолидация данных
- › При сборке DDS / ODS / Raw - проверки отставания
- › При сборке витрин - сверки

Перерыв

Теперь про практику

Полнота набора данных

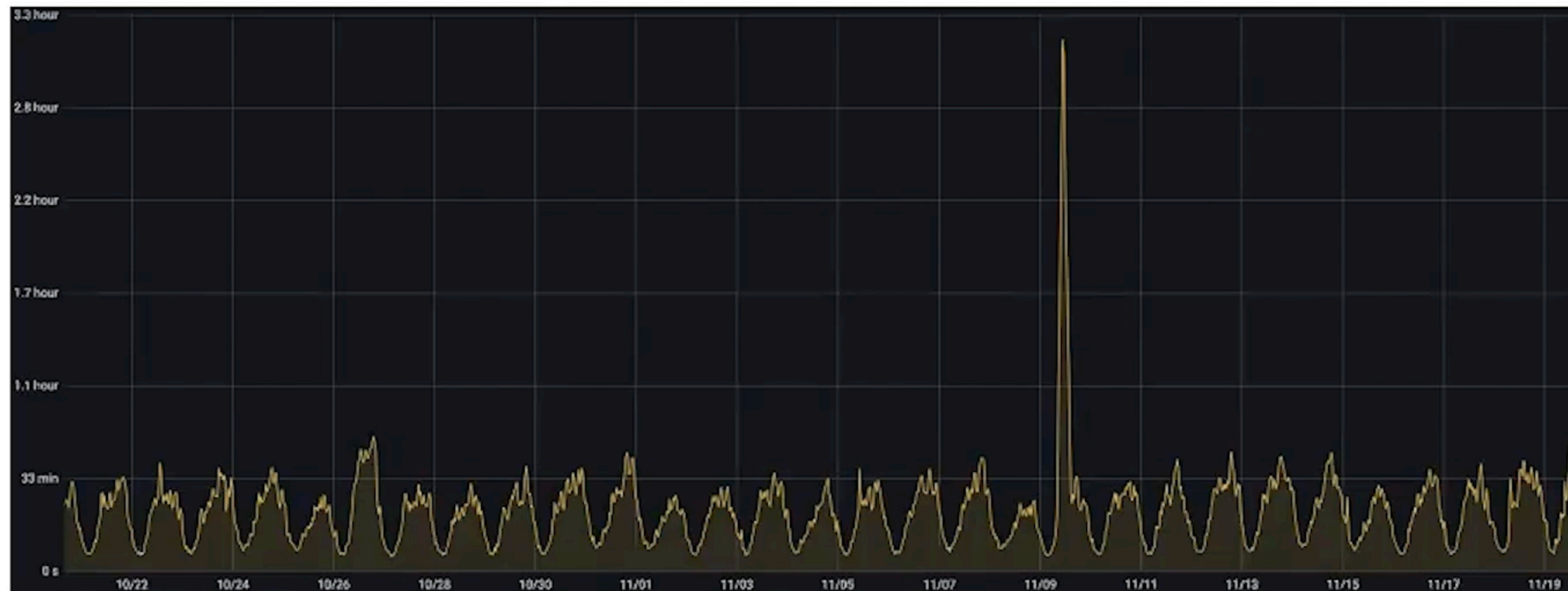
- › По совпадению множеств ключей
- › По контрольным суммам - количество записей, количество ключей, суммы по полям и т.д.

Актуальность

- › Момент обновления данных в хранилище отличается от момента последнего обновления источника не более, чем на X
- › Данные за предыдущий период полностью лежат в хранилище на момент времени X

Актуальность

- › Момент обновления данных в хранилище отличается от момента последнего обновления источника не более, чем на X
- › Данные за предыдущий период полностью лежат в хранилище на момент времени X



Консистентность

- › В каждой колонке лежат данные одного типа
- › Данные в колонках соответствуют определенному интервалу или множеству
- › Данные в колонках соответствуют какому-то формату (regex)

Полнота значений данных

Для некоторых данных могут быть пропуски значений;
Где-то это нормально и обусловлено смыслом данных;
Где-то это может говорить об ошибках в данных;

- › При записи на источник не заполняется часть полей
- › Данные теряются при записи в хранилище
- › Данные теряются при нормализации
- › В результате join склеиваются не все строки