Проектирование интеграционных решений

Лекция 16 (32)

Технологии хранилищ данных

Овчинников П.Е.

МГТУ «СТАНКИН»,

ст.преподаватель кафедры ИС

12.04.2021

Терминология: хранилища данных

ГОСТ Р ИСО/МЭК 29155-1-2016 Системная и программная инженерия. Структура сопоставительного анализа эффективности выполнения проектов информационных технологий. Часть 1. Понятия и определения

хранилище данных (repository):

организованная база данных длительного хранения, которая обеспечивает поиск и извлечение данных

Храни́лище да́нных (англ. *Data Warehouse*)

предметно-ориентированная информационная база данных, специально разработанная и предназначенная для подготовки отчётов и бизнес-анализа с целью поддержки принятия решений в организации.

Строится на базе систем управления базами данных и систем поддержки принятия решений. Данные, поступающие в хранилище данных, как правило, доступны только для чтения

Терминология: качество данных

ГОСТ Р ИСО 8000-2-2014 Качество данных. Часть 2. Словарь

метаданные (metadata): Данные, которые описывают и определяют другие данные

данные (data): Символическое представление чего-либо, частично зависящего в своем значении от метаданных

точность данных (data accuracy): Точность соответствия между значением свойства и истинным значением

истинное значение (true value): Значение параметров характеристики какого-либо объекта в определенных условиях

авторитетный источник данных (authoritative data source): Владелец процесса, производящего данные

словарь данных (data dictionary): Совокупность вводимых в словарь данных, которые можно найти по идентификатору объекта

основные данные (master data): Данные, находящиеся во владении организацией и описывающие основные объекты этой организации. На эти данные следует ссылаться при составлении транзакций

ГОСТ Р ИСО 8000-2-2014

Терминология: качество данных

ГОСТ Р ИСО 8000-2-2014 Качество данных. Часть 2. Словарь

метаданные (metadata): Данные, которые описывают и определяют другие данные

данные (data): Символическое представление чего-либо, частично зависящего в своем значении от метаданных

точность данных (data accuracy): Точность соответствия между значением свойства и истинным значением

истинное значение (true value): Значение параметров характеристики какого-либо объекта в определенных условиях

авторитетный источник данных (authoritative data source): Владелец процесса, производящего данные

словарь данных (data dictionary): Совокупность вводимых в словарь данных, которые можно найти по идентификатору объекта

основные данные (master data): Данные, находящиеся во владении организацией и описывающие основные объекты этой организации. На эти данные следует ссылаться при составлении транзакций

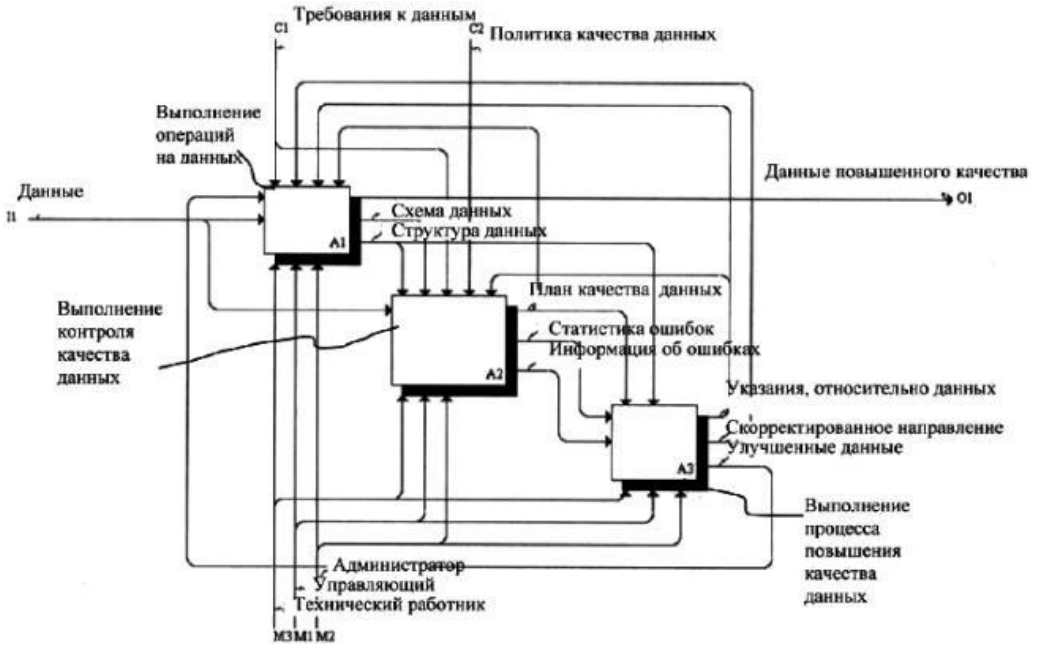
ГОСТ Р ИСО 8000-2-2014

Терминология: нормализация НСИ

Нормализация нормативно-справочной информации представляет собой приведение к стандартному виду всех данных, содержащихся в справочниках

В процессе нормализации первоначальная информация в справочниках **разбирается** и **структурируется** в соответствии с созданными правилами, одновременно выполняется **множественная классификация** элементов справочников

ГОСТ Р 56215-2014/ISO/TS 8000-150:2011 Качество данных. Часть 150. Основные данные. Структура управления качеством



Пример: нормализация адресов

Для использования «КЛАДР» - Классификатор адресов Российской Федерации на сайте, мы получаем актуальные данные Государственного реестра адресов ФНС России.

Актуальность базы: **2019.03.21**

Адрес: Москва Город -> Вадковский Переулок

Код КЛАДР: ***77000000000053400***

Код региона	Почтовый индекс	Код окато	Код налоговой
77	127055	45286585000	7707

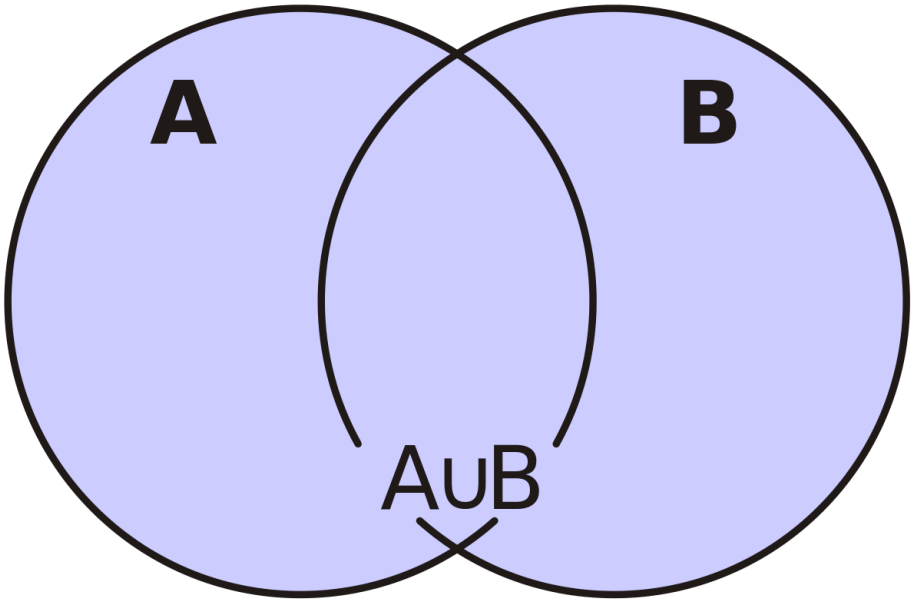
Интервал домов	Почтовый индекс	Код окато	Код налоговой
1,10стр1,10стр13,10стр2,12,16,18стр1	127055	45286585000	7707
18стр10,18стр1А,18стр4,18стр5,18стр7	127055	45286585000	7707
18стр8,18стр9,20стр1,20стр2,24/35стр1,3	127055	45286585000	7707

КЛАДР

Терминология: дедупликация

Объедине́ние мно́жеств (тж. **су́мма** или **соедине́ние**) в теории множеств — множество, содержащее в себе все элементы исходных множеств

Объединение двух множеств {A} и {B} обычно обозначается {A} ∪ {B}, но иногда можно встретить запись в виде суммы {A+B}



Объединение множеств – потенциальный источник дубликатов!

Сложение (объединение) множеств

Терминология: дедупликация

В языке SQL операция **UNION** применяется для объединения двух наборов строк, возвращаемых SQL-запросами

Оба запроса должны возвращать одинаковое число столбцов, и столбцы с одинаковым порядковым номером должны иметь совместимые типы данных

Результат получает структуру (названия и типы столбцов) первого (левого) запроса, то есть операция не является симметричной

```
(SELECT * FROM sales2005)
UNION
(SELECT * FROM sales2006);
```

Без дупликатов

```
(SELECT * FROM sales2005)
UNION ALL
(SELECT * FROM sales2006);
```

С дупликатами

Оператор UNION (Википедия)

Терминология: очистка данных

Очистка данных (англ. *Data cleansing*) — процесс выявления и исправления ошибок, несоответствий данных с целью улучшения их качества, иногда классифицируется как составная часть интеллектуального анализа данных

Очистка данных выполняется с определенными наборами данных в базах данных или файлах

Необходимость в очистке данных чаще всего возникает при интеграции различных информационных систем (хранилища данных, системы управления ресурсами предприятия, системы управления взаимодействием с клиентами)

Источники данных в различных системах часто находятся в разрозненном виде и в различных состояниях. Преобразования выполняются автоматически (в соответствии с набором правил) либо вручную (в интерактивном режиме).

Наиболее типичные предметные области, подлежащие очистке и исправлению в корпоративных информационных системах — сведения о лицах и организациях, адресная и контактная информация, также подлежит очистке любая справочная информация, вносимая вручную в текстовом виде.

Очистка данных (Википедия)

Терминология: обогащение данных

Процесс насыщения данных новой информацией, которая позволяет сделать их более ценными и значимыми с точки зрения решения той или иной аналитической задачи

Существует два основных метода обогащения данных - **внешнее** и **внутреннее**

Внешнее обогащение предполагает привлечение дополнительной информации из источников, которые находятся вне информационной системы предприятия

Практически источником информации для обогащения данных могут быть любые организации, которые в процессе своей деятельности собирают, структурируют и хранят сведения, связанные с их деятельностью

Внутреннее обогащение не предполагает привлечения какой-либо внешней информации. Оно обычно связано с получением и включением в набор данных полезной информации, которая отсутствует в явном виде, но может быть тем или иным способом получена с помощью манипуляций с имеющимися данными.

Затем, эта информация встраивается в виде новых полей или даже таблиц в хранилище данных и может быть использована для дальнейшего анализа

Обогащение данных (Мнение)

НСИ, MDM

http://solutions.1c.ru/catalog/mdm/features

НСИ, MDM

http://solutions.1c.ru/catalog/mdm/features

Терминология:ETL

ETL (от англ. *Extract, Transform, Load* — дословно «извлечение, преобразование, загрузка») — один из основных процессов в управлении хранилищами данных, который включает в себя:

- извлечение данных из внешних источников;
- трансформацию и очистку данных
- загрузку в хранилище данных

С точки зрения процесса ETL, архитектуру хранилища данных можно представить в виде трёх компонентов:

- источник данных: содержит структурированные данные в виде таблиц, совокупности таблиц или просто файла (данные в котором разделены символами-разделителями);
- промежуточная область: содержит вспомогательные таблицы, создаваемые временно, и, исключительно для организации процесса выгрузки.
- получатель данных: хранилище данных или база данных, в которую должны быть помещены извлечённые данные.

Перемещение данных от источника к получателю называют потоком данных

ETL (Википедия)

Терминология:ETL

Извлечение данных в ETL

Начальным этапом процесса ETL является процедура извлечения записи из источников данных и подготовка их к процессу преобразования

При разработке процедуры извлечения данных, в первую очередь необходимо определить частоту выгрузки данных из OLTP-систем или отдельных источников

Выгрузка данных занимает определённое время, которое называется окном выгрузки.

Процедуру извлечения данных можно реализовать двумя способами:

- извлечение данных с помощью специализированных программных средств;
- извлечение данных средствами той системы, в которой они хранятся.

После извлечения данные помещаются в так называемую «промежуточную область», где для каждого источника данных создаётся своя таблица или отдельный файл, или и то и другое

ETL (Википедия)

Терминология:ETL

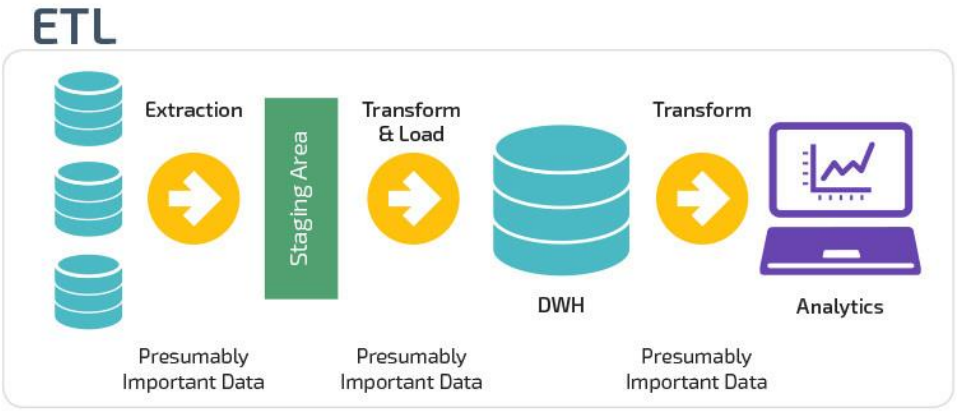
Преобразование данных

Цель этого этапа — подготовка данных к размещению в хранилище данных и приведение их к виду более удобному для последующего анализа

При этом должны учитываться некоторые, выдвигаемые аналитиком, требования, в частности, к уровню качества данных. Поэтому в процессе преобразования может быть задействован самый разнообразный инструментарий, начиная с простейших средств ручного редактирования данных и заканчивая системами, реализующими сложные методы обработки и очистки данных

В процессе преобразования данных в рамках ETL чаще всего выполняются следующие операции:

- преобразование структуры данных
- агрегирование данных
- перевод значений
- создание новых данных
- очистка данных



ETL (Википедия)

Терминология:ETL

Загрузка данных

Процесс загрузки заключается в переносе данных из промежуточных таблиц в структуру хранилища данных. При очередной загрузке в хранилище данных переносится не вся информация из источников, а только та, которая была изменена в течение промежуточного времени, прошедшего с предыдущей загрузки

При этом выделяют два потока:

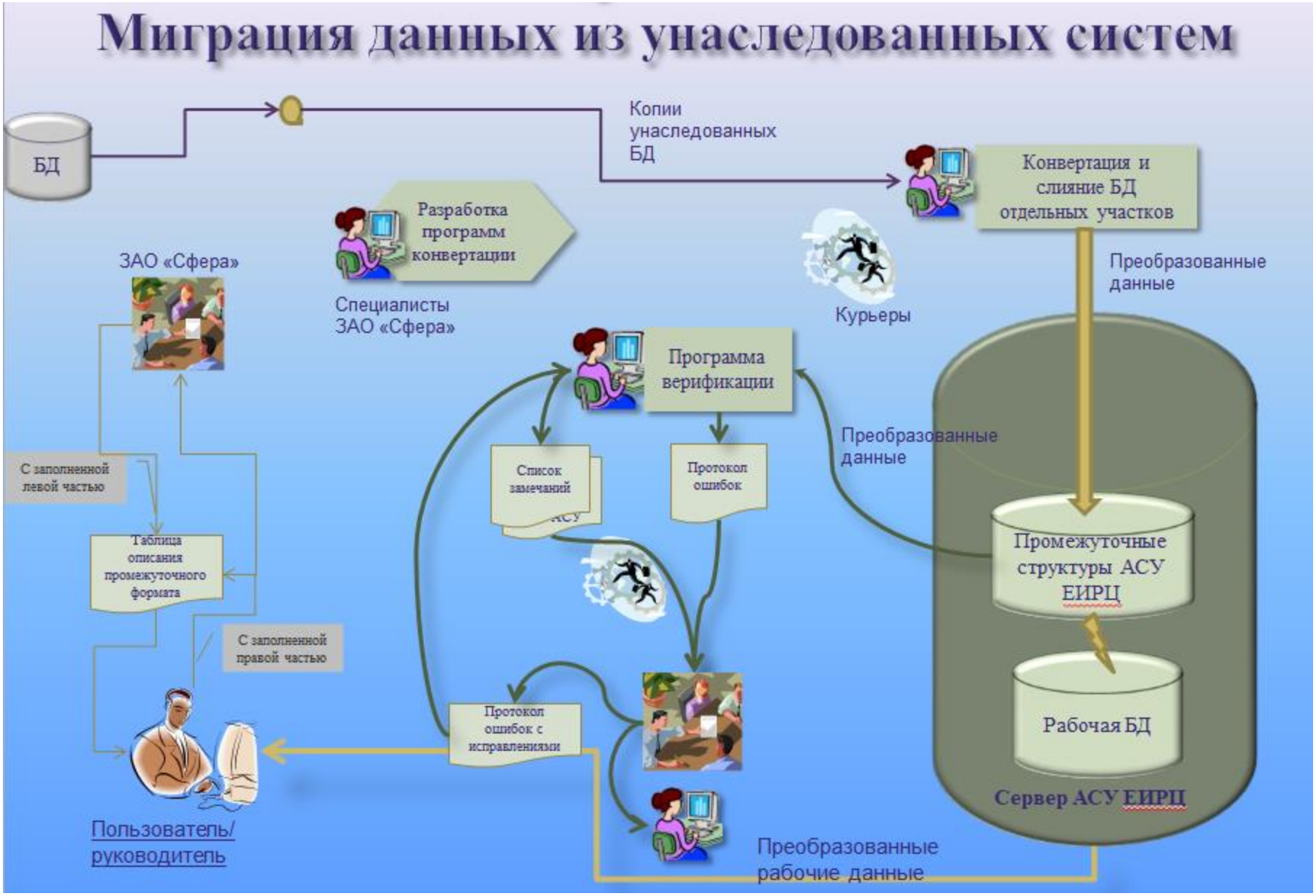
- **поток добавления** — в хранилище данных передается новая, ранее не существовавшая информация;
- **поток обновления** (дополнения) — в хранилище данных передается информация, которая существовала ранее, но была изменена или дополнена.

Для распределения загружаемых данных на потоке используются средства данных. Они фиксируют состояние данных в некоторые моменты времени и определяют, какие данные были изменены или дополнены

ETL (Википедия)

Терминология: миграция данных

Терминология: миграция данных



Терминология: большие данные

Больши́е да́нные (англ. *big data*, [ˈbɪg ˈdeɪtə]) — обозначение структурированных и неструктурированных данных огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами и альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence

В качестве определяющих характеристик для больших данных традиционно выделяют «**три V**»:

- **объём** (англ. *volume*, в смысле величины физического объёма),
- **скорость** (*velocity* в смыслах как скорости прироста, так и необходимости высокоскоростной обработки и получения результатов),
- **многообразие** (*variety*, в смысле возможности одновременной обработки различных типов структурированных и полуструктурированных данных)

в дальнейшем возникли различные вариации и интерпретации этого признака

Большие данные (Википедия)

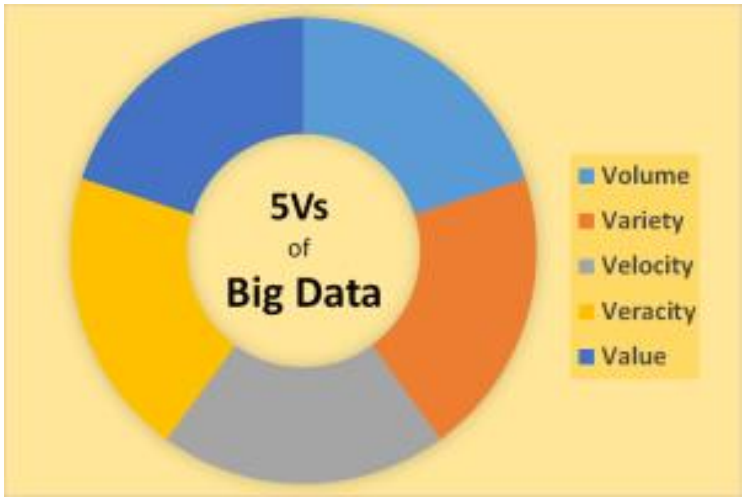
Терминология: большие данные

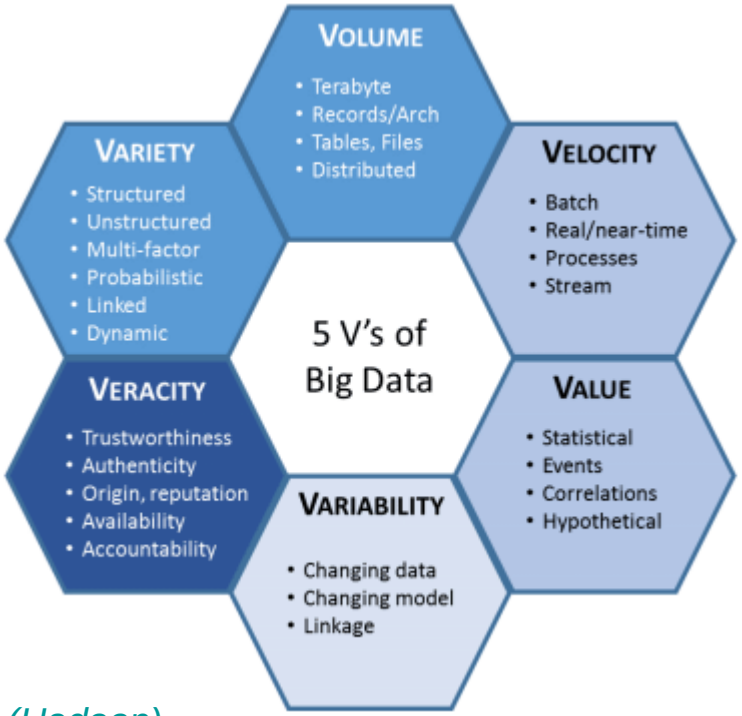
4. Правдивость (Veracity)

Точки данных, которые были собраны и сохранены из различных источников в различных формах, часто имеют дело с неточностью. При этом нам приходится иметь дело с низким качеством данных, в том числе в огромных объемах (например, в сообщениях Twitter с хеш-тегами, опечатками, сокращениями и разговорной речью), которые не являются точными и неопределенными

4V (IBM)

Терминология: большие данные





5. Ценность (Value)
Независимо от того, являются ли данные большими или небольшими, независимо от того, были ли они получены в любом месте и в каком бы то ни было формате, они должны иметь определенную ценность - это означает, что мы можем правильно использовать данные по их правильной причине для их достоверности

Значение, ценность или функциональность данных для тех, кто их потребляет, по-видимому, наиболее важны для различных фирм или организаций. Кроме того, мы знаем, что данные сами по себе не имеют никакого значения или полезности, но все же нам нужны ценные данные для получения информации.

5V (Hadoop)

Пример: технология MapReduce

MapReduce — это фреймворк для вычисления некоторых наборов распределенных задач с использованием большого количества компьютеров (называемых «нодами»), образующих кластер

Работа MapReduce состоит из двух шагов: Map и Reduce, названных так по аналогии с одноименными функциями высшего порядка, map и reduce

На Map-шаге происходит предварительная обработка входных данных. Для этого один из компьютеров (называемый главным узлом — master node):

- **получает** входные данные задачи
- **разделяет** их на части и
- **передает** другим компьютерам (рабочим узлам — worker node) для предварительной обработки

На Reduce-шаге происходит свёртка предварительно обработанных данных.

Главный узел получает ответы от рабочих узлов и на их основе формирует результат — решение задачи, которая изначально формулировалась.

MapReduce (Википедия)

Пример: технологии кластерного анализа

Кластерный анализ (англ. *cluster analysis*) — многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Кластерный анализ решает следующие основные задачи:

- разработка **типологии** или классификации
- исследование полезных концептуальных **схем** группирования объектов
- порождение **гипотез** на основе исследования данных
- проверка гипотез или **исследования** для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных

Применение кластерного анализа предполагает следующие этапы:

- отбор **выборки** для кластеризации
- определение множества переменных, по которым будут оцениваться объекты в выборке, то есть **признакового пространства**
- вычисление значений той или иной **меры сходства** (или различия) между объектами
- применение метода кластерного анализа для **создания групп** сходных объектов
- проверка **достоверности** результатов кластерного решения

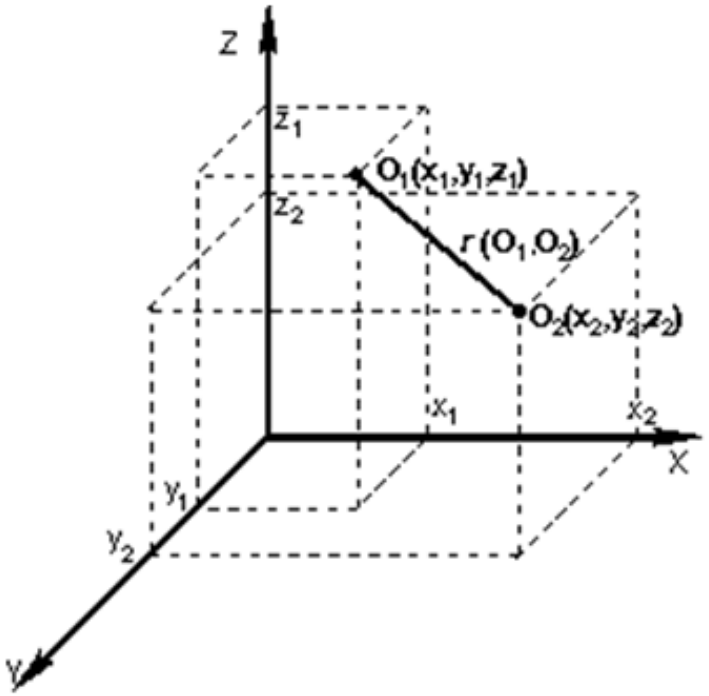
Кластерный анализ (Википедия)

Пример: евклидово пространство

Евкли́дово простра́нство (также **эвкли́дово простра́нство**) — в изначальном смысле, пространство, свойства которого описываются аксиомами евклидовой геометрии

Евклидово расстояние между точками находится по формуле

$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, где n – количество измерений



Расстояние между двумя точками в пространстве трех измерений

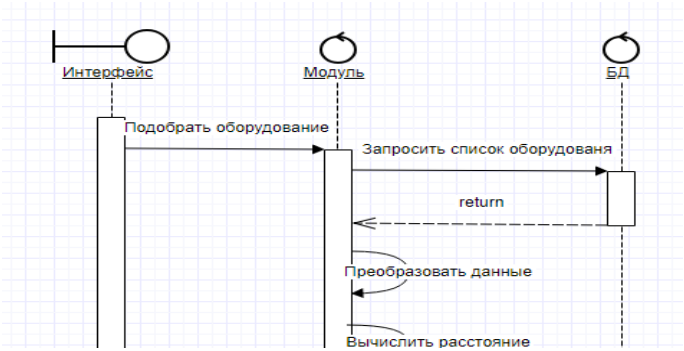
Эвклидово пространство (Википедия)

Пример: расчет расстояний

1. Получаются значения эталона, заданные пользователем, по которым будет выполняться сравнение

2. Из справочника получается список элементов с указанием названия, параметров и их значений

3. Так как параметры могут иметь разные размерности, то для каждого элемента они приводятся к единому виду, где единица – величина, равная значениям эталона.



V1 (эталон)	V2	Расстояние
Число	Пусто	1
Число	Число	$\sqrt[2]{(V1-V2)^2}$
Число	Строка	1
Число	Диапазон	Берется среднее значение диапазона и считается, как «число-число»
Диапазон	Число	
Диапазон	Диапазон	
Строка	Пусто	1
Строка	Число	1
Строка	Строка	0 – совпадают, 1 – не совпадают
Строка	Диапазон	1
Диапазон	Пусто	1
Диапазон	Строка	1

Эвклидово пространство (Википедия)