Проектирование высоконагруженных и аналитических систем

Лекция 18 (34)

Аналитические и экспертные системы

Овчинников П.Е.

МГТУ «СТАНКИН»,

ст.преподаватель кафедры ИС

Терминология: OLTP-системы

Транза́кция (англ. *transaction*) — группа последовательных операций с базой данных, которая представляет собой логическую единицу работы с данными

Транзакция может быть выполнена либо целиком и успешно, соблюдая целостность данных и независимо от параллельно идущих других транзакций, либо не выполнена вообще, и тогда она не должна произвести никакого эффекта

OLTP (англ. *Online Transaction Processing*), **транзакционная система** — обработка транзакций в реальном времени.

Способ организации базы данных, при котором система работает с **небольшими** по размерам транзакциями, но идущими **большим потоком**, и при этом клиенту требуется от системы **минимальное время отклика**

OLTP-системы предназначены для:

- **ввода**
- структурированного **хранения**
- **обработки**

информации (операций, документов) в режиме реального времени

Терминология: OLTP-системы

OLTP-приложениями охватывается широкий спектр задач во многих отраслях — автоматизированные банковские системы, ERP-системы (системы планирования ресурсов предприятия), банковские и биржевые операции, в промышленности — регистрация прохождения детали на конвейере, фиксация в статистике посещений очередного посетителя веб-сайта, автоматизация бухгалтерского, складского учёта и учёта документов и т. п.

Приложения OLTP, как правило, автоматизируют структурированные, **повторяющиеся задачи обработки** данных, такие как ввод заказов и банковские транзакции. OLTP-системы проектируются, настраиваются и оптимизируются для выполнения максимального количества транзакций за короткие промежутки времени. Как правило, большой гибкости здесь не требуется, и чаще всего используется фиксированный набор надёжных и безопасных методов ввода, модификации, удаления данных и выпуска оперативной отчётности

Показателем эффективности является количество **транзакций**, выполняемых **за секунду**. Обычно аналитические возможности OLTP-систем сильно ограничены (либо вообще отсутствуют).

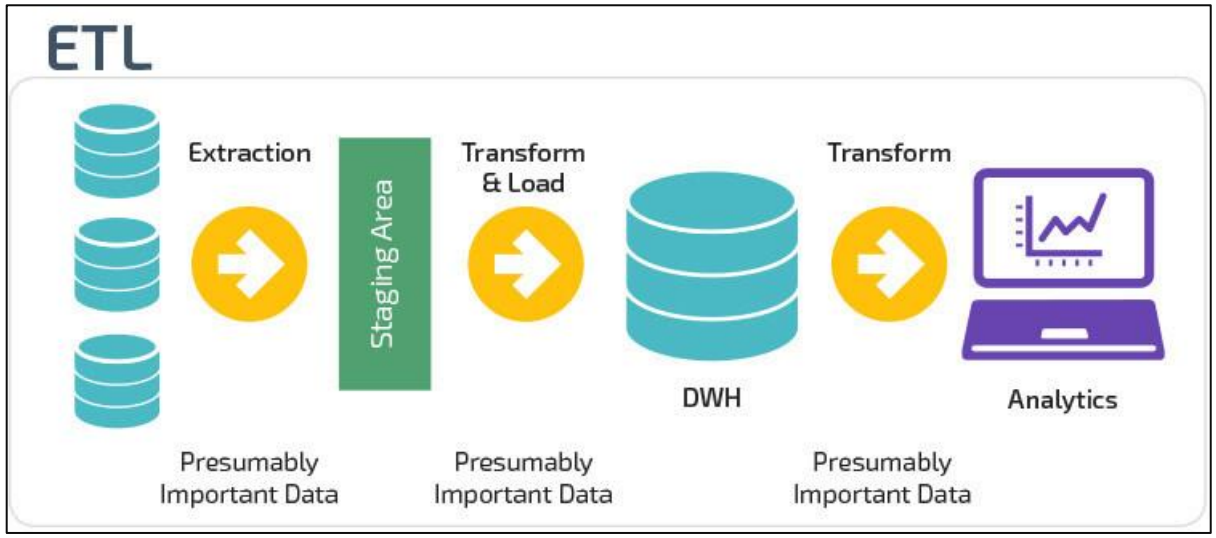
Терминология: OLAP-системы

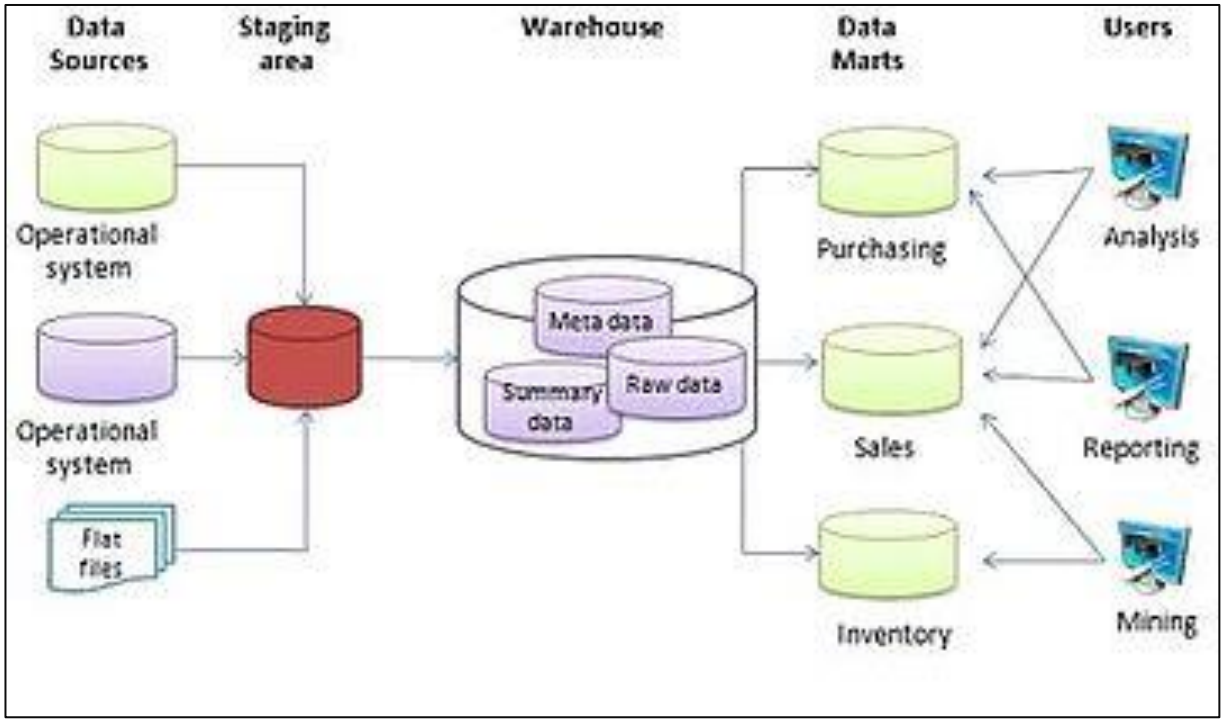
OLAP (англ. *online analytical processing*, аналитическая обработка в реальном времени) — технология обработки данных, заключающаяся в подготовке **суммарной (агрегированной) информации** на основе больших массивов данных, структурированных по многомерному принципу.

OLTP vs. OLAP Differences

	OLTP	OLAP
Organization	By workflow per application	By dimension and business subject
Data Retention	Short term (2-6 months)	Long term (2-5 years)
Data Integration	Minimal or none	High, as part of ETL process
Data Storage	Gigabytes	Terabytes
Use	Real time Write & update Evenly distributed usage Transactional data	Batch load Reporting, read-only Spiked usage (based on time of warehouse loads)

Терминология: OLAP-системы





Терминология: OLAP-системы

Реализации технологии OLAP являются компонентами программных решений класса Business Intelligence

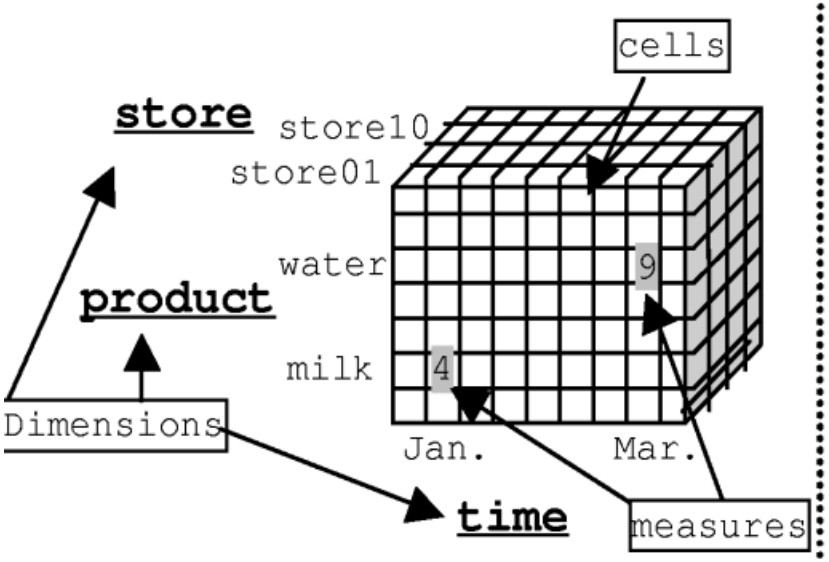
Многомерное моделирование является методом моделирования и визуализации данных как множества числовых или лингвистических показателей или параметров (measures), которые описывают общие аспекты деятельности организации

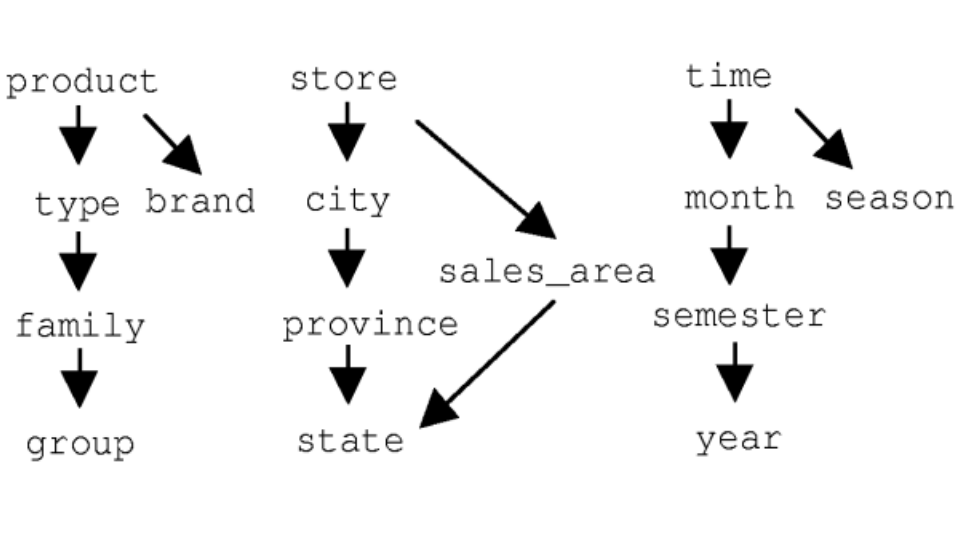
Метод многомерного моделирования базируется на следующих основных понятиях:

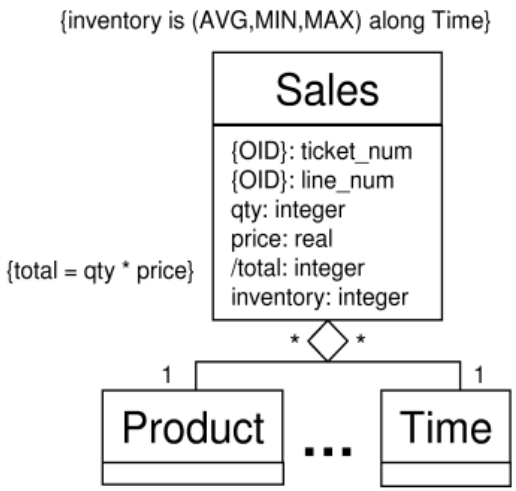
- **Факт (Fact)** — набор связанных элементов данных, содержащих метрики и описательные данные.
- **Атрибут (Attribute)** – описание характеристики реального объекта предметной области.
- **Измерение (Dimension)** — интерпретация факта с некоторой точки зрения в реальном мире.
- **Параметр, метрика или показатель (Measure)** — числовая характеристика факта
- **Гранулированность (Granularity)** – уровень детализации данных.

Dimensional modeling *Многомерное моделирование (Интуит)*

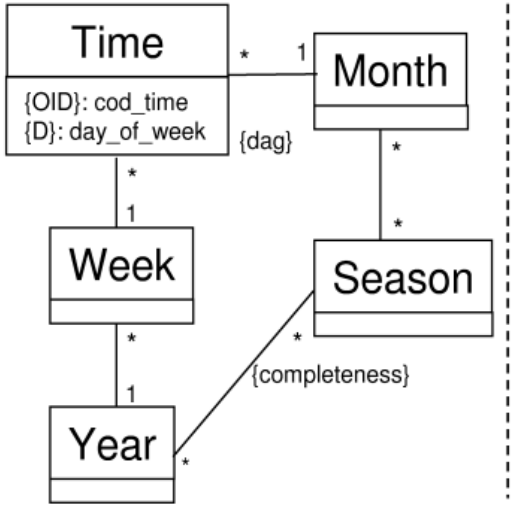
Терминология: OLAP-куб



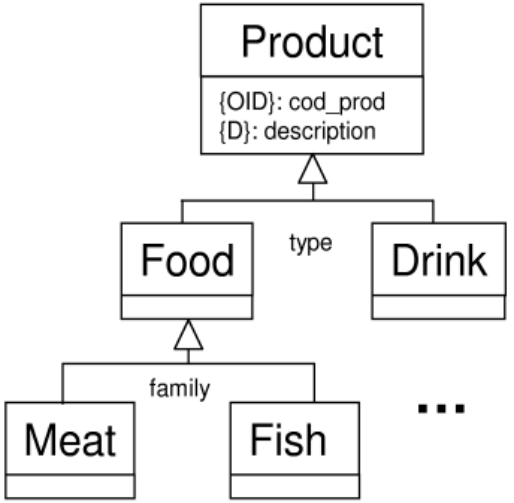




(a)

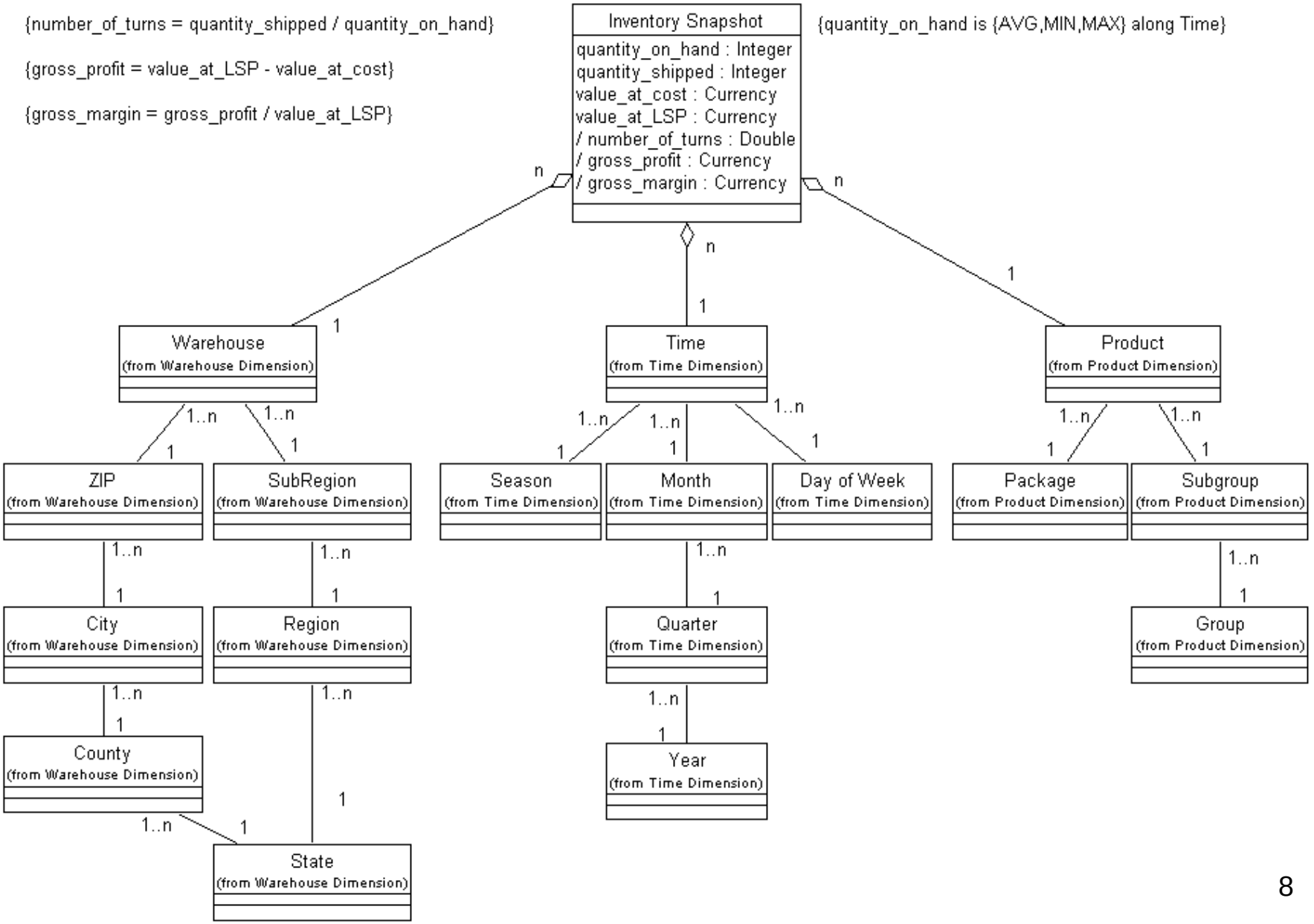


(b)

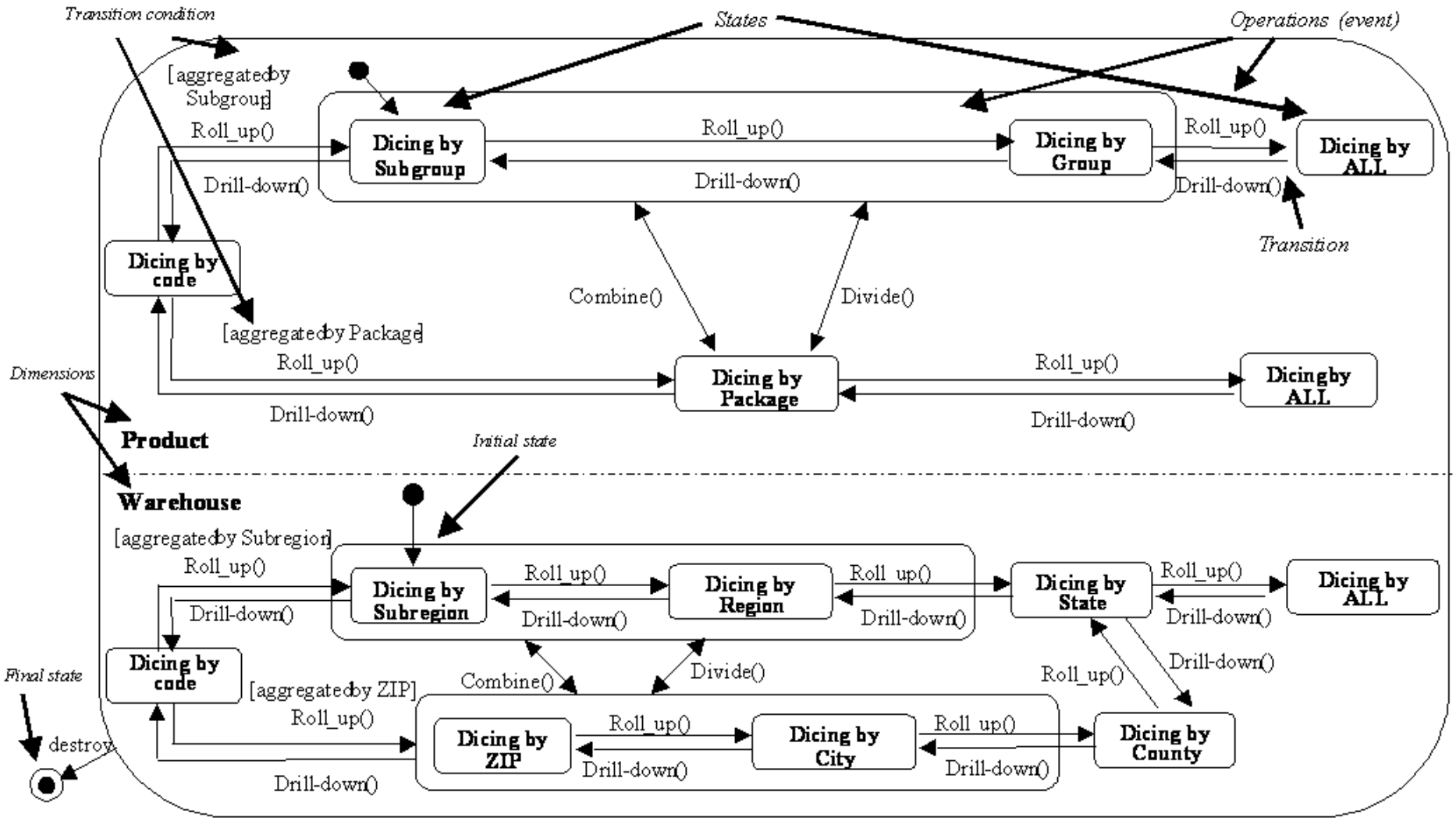


(c)

Терминология: измерение



Терминология: roll up и drill down



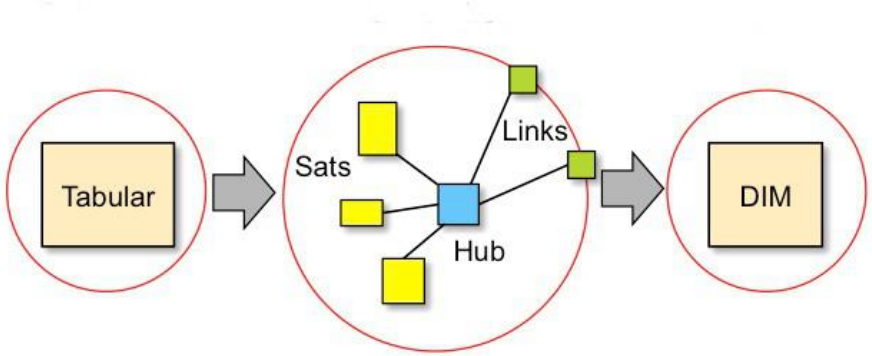
[UML Диаграмма состояний]

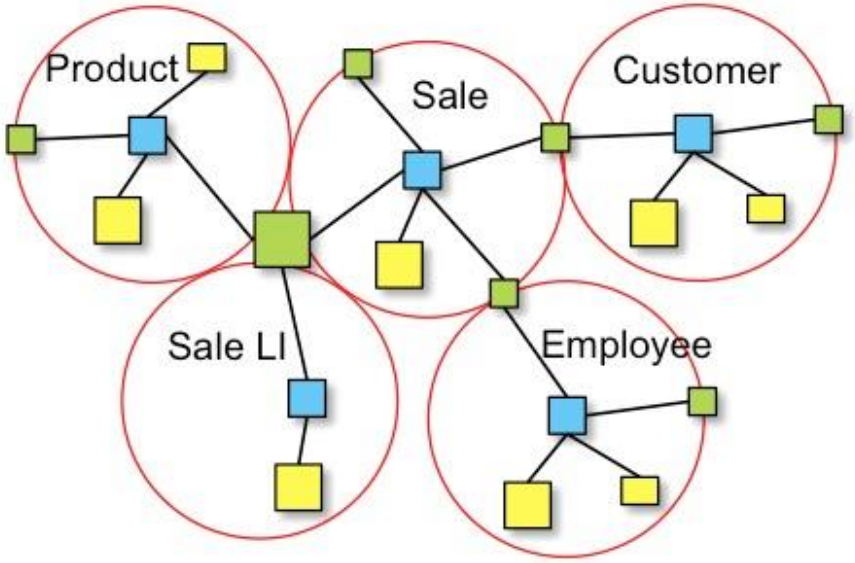
Нормальные формы: Data Vault

Модель Data Vault представляет собой набор связанных между собой нормализованных таблиц, ориентированных на хранение детализированной информации с возможностью отслеживания происхождения данных и поддерживающих одну или несколько областей бизнеса.

В модели Data Vault используется всего **три типа таблиц**:

- **Hub** обеспечивает представление функциональных областей предметной области
- **Link** обеспечивает транзакционную связь между Hub-таблицами
- **Satellite** предоставляет детализацию первичного ключа Hub-таблицы

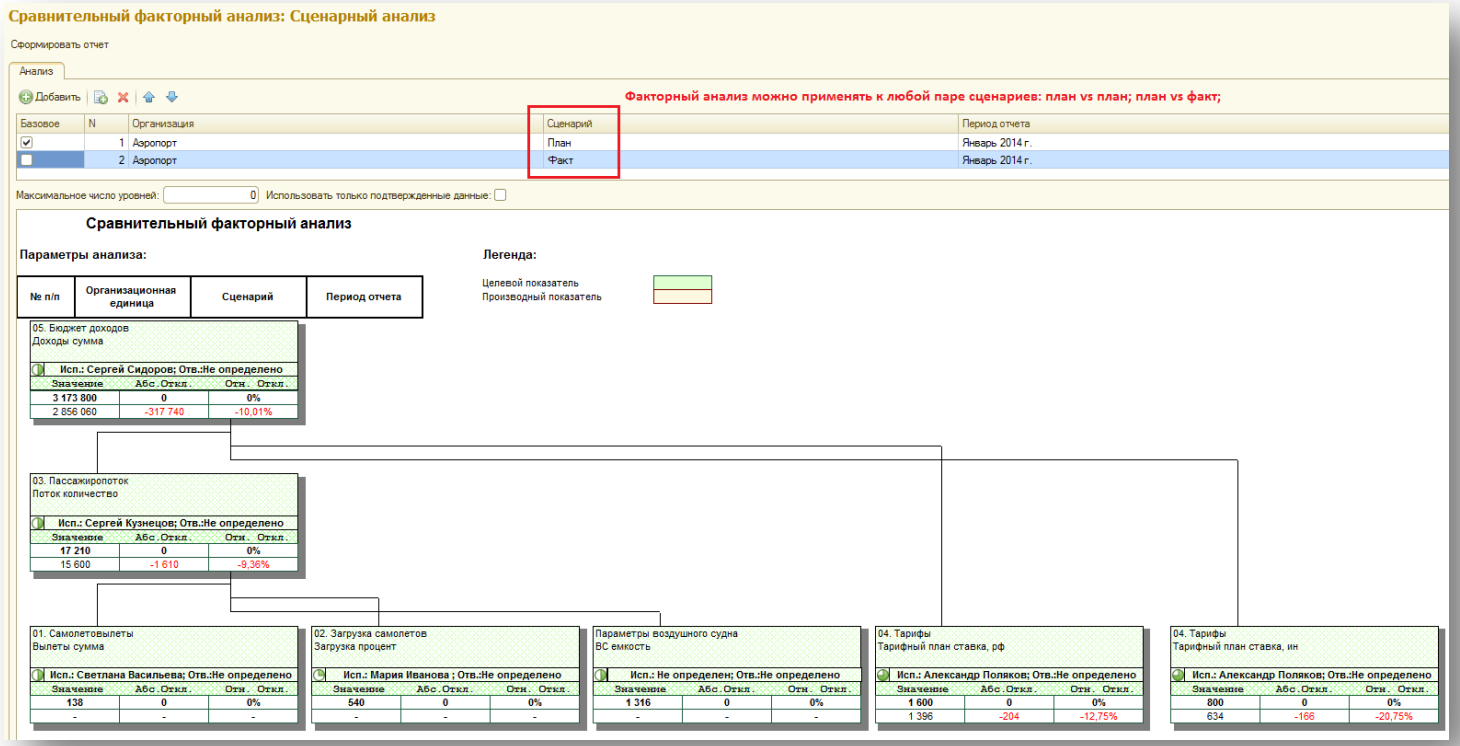




[Data Vault](#)
[Модель Свод данных (Интуит)](#)

Терминология: CPM-системы

Управление эффективностью деятельности организации (английские термины CPM, BPM, EPM) — это набор управленческих процессов (планирования, организации выполнения, контроля и анализа), которые позволяют бизнесу определить стратегические цели и затем оценивать и управлять деятельностью по достижению поставленных целей при оптимальном использовании имеющихся ресурсов. Это система управления, построенная на принципах управления стоимостью бизнеса.



Терминология: консолидация

Консолидация данных

комплекс методов и процедур, направленных на **извлечение** данных из различных источников, **обеспечение** необходимого уровня их информативности и **качества**, преобразование в **единый формат**, в котором они могут быть загружены в хранилище данных или аналитическую систему

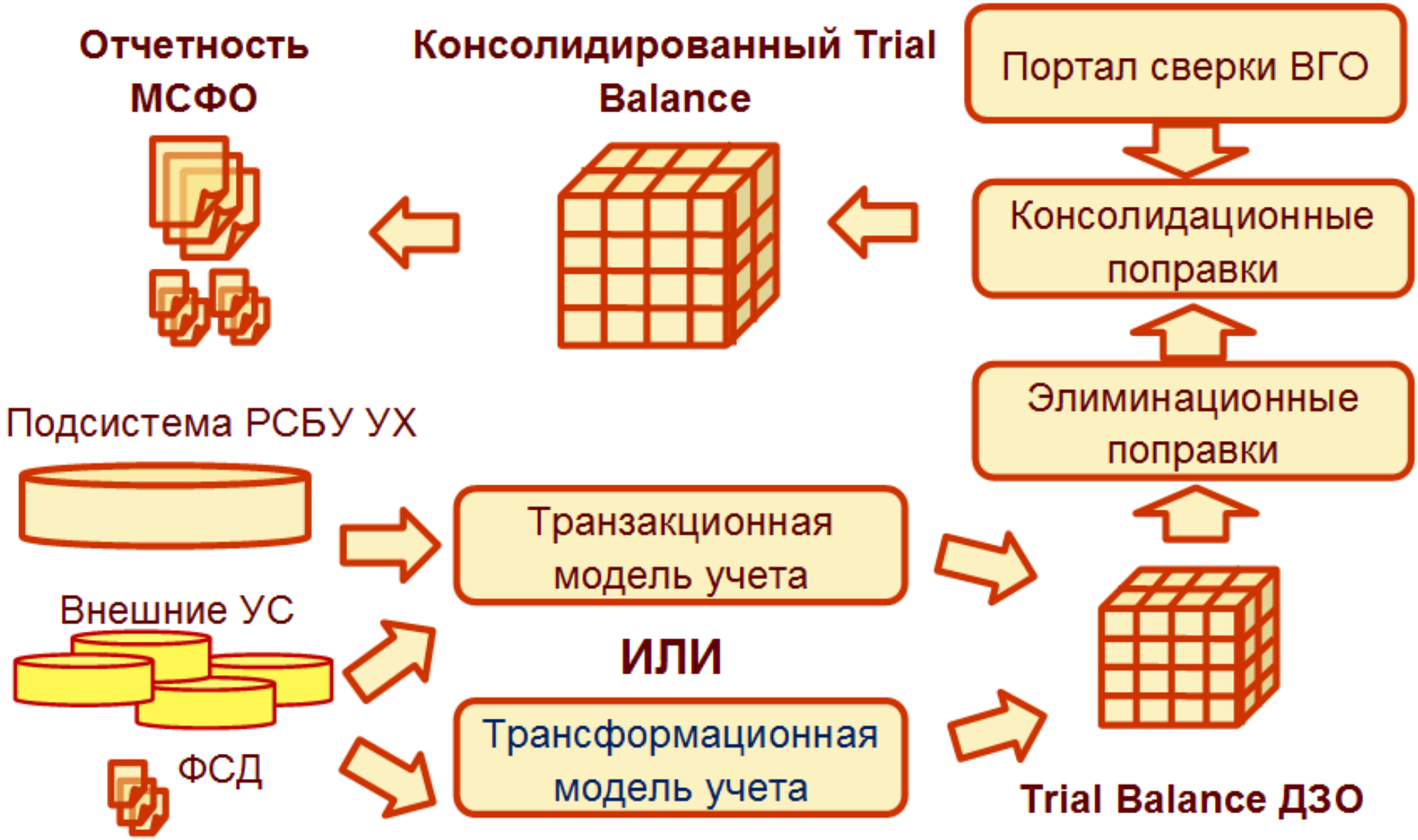
Консолидированная (сводная) отчетность - это система показателей, отражающих финансовое положение и финансовые результаты предприятия по истечении отчетного периода и включающая данные о зависимых обществах, являющихся юридическими лицами по законодательству места их государственной регистрации

Трансформация бухгалтерской (финансовой) отчетности - это процесс составления отчетности в соответствии с МСФО путем перегруппировки учетной информации и корректировки статей отчетности, подготовленной по правилам российской системы бухгалтерского учета

Элиминация - это процесс исключение внутригрупповых операций при консолидации отчётности

Консолидация (1С)

Терминология: консолидация



Консолидация (1С)

Терминология: экспертные системы

Эксперт́ная систе́ма (англ. *expert system*)

компьютерная система, способная частично заменить специалиста-эксперта в разрешении проблемной ситуации.

Современные экспертные системы начали разрабатываться исследователями искусственного интеллекта в 1970-х годах, а в 1980-х годах получили коммерческое подкрепление

Важнейшей частью экспертной системы являются базы знаний как модели поведения экспертов в определённой области знаний с использованием процедур логического вывода и принятия решений, иными словами, базы знаний — совокупность **фактов** и **правил логического вывода** в выбранной предметной области деятельности

Экспертная система анализирует ситуацию и даёт рекомендации по разрешению проблемы

Как правило, база знаний экспертной системы содержит факты (статические сведения о предметной области) и правила — набор инструкций, применяя которые к известным фактам можно получать новые факты

Экспертные системы (Википедия)

Терминология: СППР

Систе́ма подде́ржки приня́тия реше́ний (СППР) (англ. *Decision Support System, DSS*)

> компьютерная автоматизированная система, целью которой является помощь людям, принимающим решение в сложных условиях для полного и объективного анализа предметной деятельности

СППР возникли в результате слияния управленческих информационных систем и систем управления базами данных и используют различные методы:

- информационный поиск,
- интеллектуальный анализ данных,
- поиск знаний в базах данных,
- рассуждение на основе прецедентов,
- имитационное моделирование,
- эволюционные вычисления и генетические алгоритмы,
- нейронные сети,
- ситуационный анализ,
- когнитивное моделирование и др.

Некоторые из этих методов были разработаны в рамках искусственного интеллекта

СППР (Википедия)

Терминология: машинное обучение

Машинное обучение (англ. Machine Learning)
обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться

Различают два типа обучения:

- ***Обучение по прецедентам***, или ***индуктивное обучение***, основано на выявлении общих закономерностей по частным эмпирическим данным
- ***Дедуктивное обучение*** предполагает формализацию знаний экспертов и их перенос в компьютер в виде базы знаний

Дедуктивное обучение принято относить к области экспертных систем, поэтому термины *машинное обучение* и *обучение по прецедентам* можно считать синонимами

Обучение с учителем (supervised learning) — наиболее распространённый случай. Каждый прецедент представляет собой пару «объект, ответ». Требуется найти функциональную зависимость ответов от описаний объектов и построить алгоритм, принимающий на входе описание объекта и выдающий на выходе ответ. Функционал качества обычно определяется как средняя ошибка ответов, выданных алгоритмом, по всем объектам выборки.

Терминология: машинное обучение

Задача классификации (classification) отличается тем, что множество допустимых ответов конечно. Их называют метками классов (class label). Класс — это множество всех объектов с данным значением метки

Задача регрессии (regression) отличается тем, что допустимым ответом является действительное число или числовой вектор.

Задача ранжирования (learning to rank) отличается тем, что ответы надо получить сразу на множестве объектов, после чего отсортировать их по значениям ответов
Может сводиться к задачам классификации или регрессии. Часто применяется в информационном поиске и анализе текстов

Задача прогнозирования (forecasting) отличается тем, что объектами являются отрезки временных рядов, обрывающиеся в тот момент, когда требуется сделать прогноз на будущее
Для решения задач прогнозирования часто удаётся приспособить методы регрессии или классификации, причём во втором случае речь идёт скорее о задачах принятия решений

Терминология: машинное обучение

Обучение без учителя (unsupervised learning). В этом случае ответы не задаются, и требуется искать зависимости между объектами.

Задача кластеризации (clustering) заключается в том, чтобы сгруппировать объекты в кластеры, используя данные о попарном сходстве объектов. Функционалы качества могут определяться по-разному, например, как отношение средних межкластерных и внутрикластерных расстояний.

Задача поиска ассоциативных правил (association rules learning). Исходные данные представляются в виде признаковых описаний. Требуется найти такие наборы признаков, и такие значения этих признаков, которые особенно часто (неслучайно часто) встречаются в признаковых описаниях объектов

Задача фильтрации выбросов (outliers detection) — обнаружение в обучающей выборке небольшого числа нетипичных объектов. В некоторых приложениях их поиск является самоцелью (например, обнаружение мошенничества). В других приложениях эти объекты являются следствием ошибок в данных или неточности модели, то есть шумом, мешающим настраивать модель, и должны быть удалены из выборки

Терминология: машинное обучение

Задача построения доверительной области (quantile estimation) — области минимального объёма с достаточно гладкой границей, содержащей заданную долю выборки

Задача сокращения размерности (dimensionality reduction) заключается в том, чтобы по исходным признакам с помощью некоторых функций преобразования перейти к наименьшему числу новых признаков, не потеряв при этом никакой существенной информации об объектах выборки

Задача заполнения пропущенных значений (missing values) — замена недостающих значений в матрице объекты–признаки их прогнозными значениями