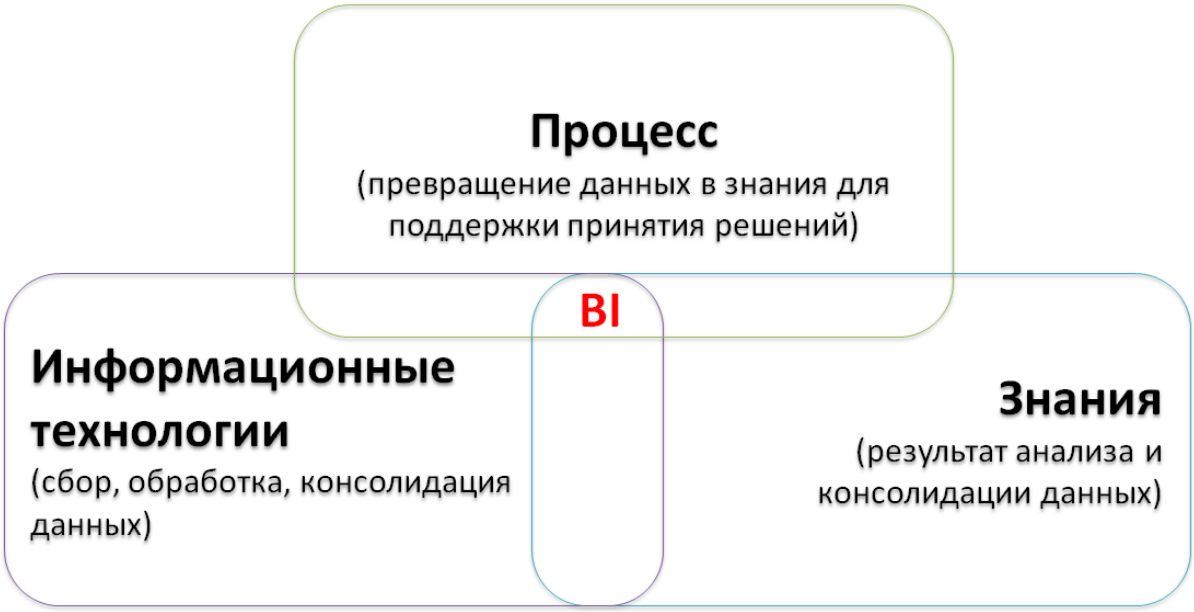
Бизнес-аналитика. Принципы Bisuness Intelligence. Системы поддержки принятия решений. Принципы построения и организации.

Что такое Business Intelligence?

Gartner: пользователецентрический процесс, включающий доступ и исследование информации, ее анализ, выработку интуиции и понимания, которые ведут к улучшенному и неформальному принятию решений



Пирамида информации

Уровень оперативной информации. На этом уровне ИТ обеспечивают работу с данными на уровне бизнес-процедур компании. Данные в автоматизированных системах являются хорошо структурированными и детальными. С этими данными работают специалисты компании: бухгалтеры, менеджеры продаж, плановики и т.д.

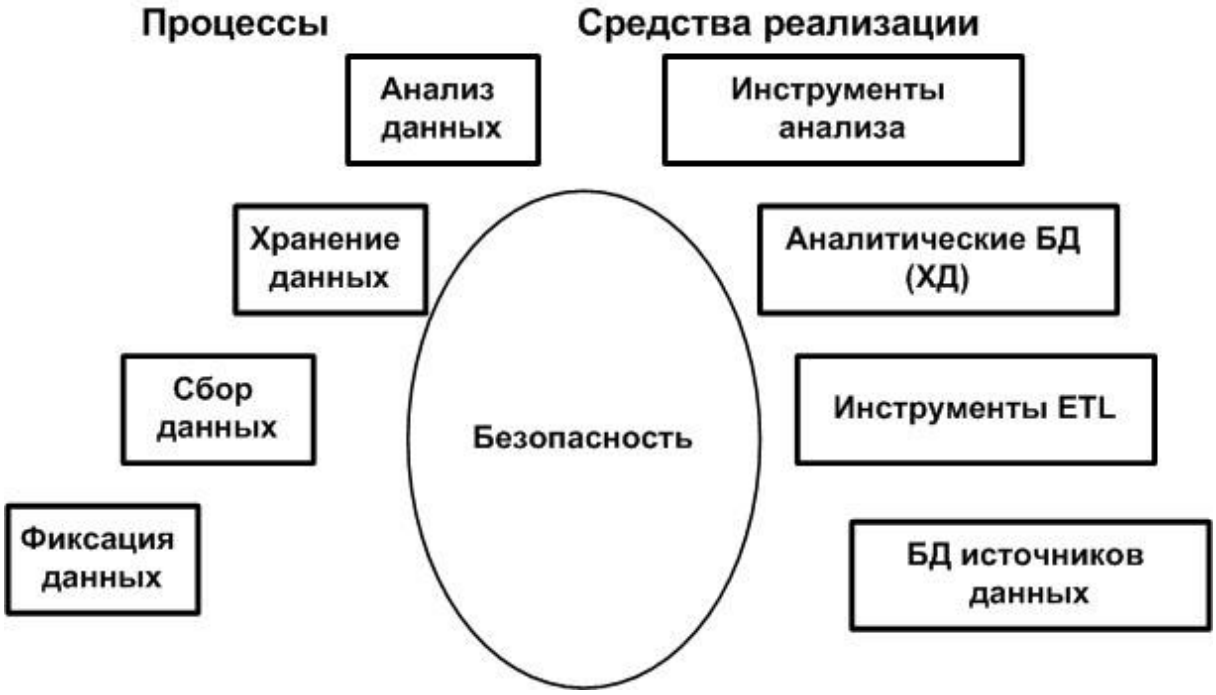
Уровень тактической информации. На этом уровне ИТ обеспечивают интеграцию данных на уровне бизнес-процессов оперативного управления производством в рамках подразделений компании. С этими данными работают руководители подразделений компании при выполнении ежедневных *производственных заданий*.

Уровень стратегической информации. На этом уровне ИТ обеспечивают интеграцию данных на уровне бизнес-процессов по направлениям хозяйственной деятельности компании. С этими данными работают аналитики и руководители высшего звена компании, которые готовят стратегические решения развития и деятельности компании на рынке.

Уровень принятия решений. На этом уровне ИТ обеспечивают интеграцию и агрегацию данных на уровне бизнес-процессов компании для руководителей высшего звена компании. Этот уровень обеспечивает информационную поддержку принятия решений

Пирамида анализа

Типовой состав BI-систем

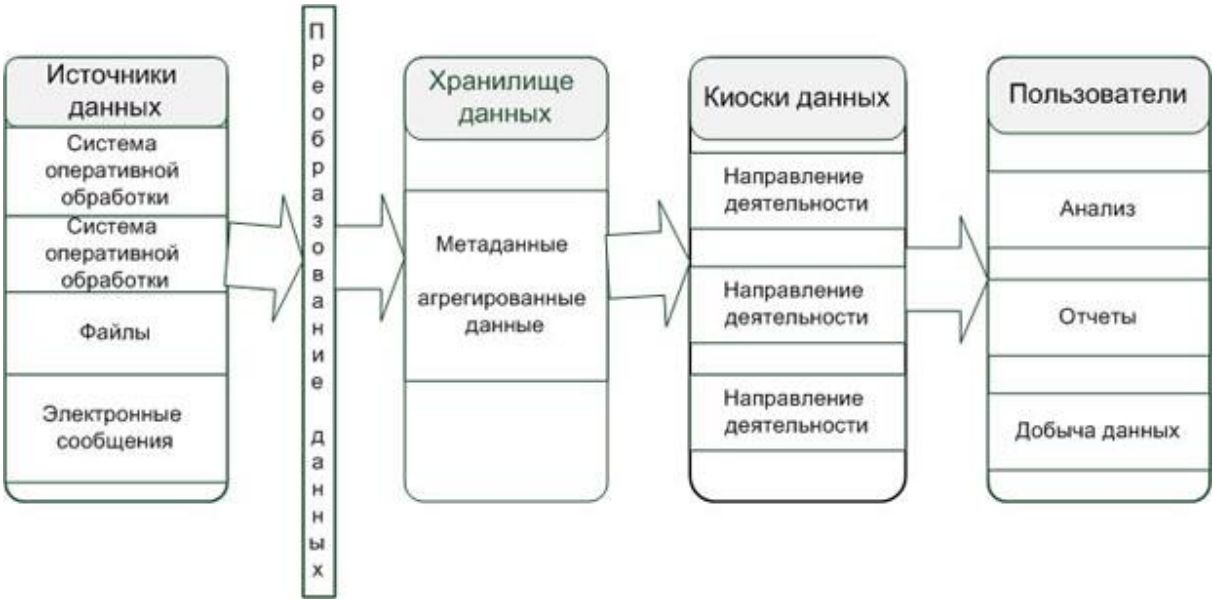


Для чего требуется выделенное хранилище данных?

В чём принципиальная разница между БД и ХД?

	Системы обработки данных	Системы хранения данных
Частота обновления	Real-time	Periodical
Цель структурирования данных	Обеспечение целостности данных	Обеспечение простоты выполнения запросов
Оптимизация данных	Для обеспечения процесса выполнения транзакций	Для обеспечения процесса выполнения выборки данных

Каким образом применяется ХД:



Метаданные. Метаданные представляют собой репозиторий, который играет роль справочника о данных. Он включает терминологию предметной области, сведения об источниках данных, описание источников исходных данных, сведения об алгоритмах обработки исходных данных и т.д.

Способы анализа

OLAP (Online Analytical Processing)

Результат (качество) анализа зависит от человека, использующего систему аналитики

Data Mining

Результат (качество) анализа зависит от математических моделей, используемых в системе

- **Изучение предметной области**
- **Создание модельных данных**: селекция данных
- **Очистка данных** и предобработка: (до 60% времени!)
- Уменьшение размерности данных и трансформации

- Выбор алгоритмов Data Mining
- Data Mining: поиск интересных паттернов
- Оценка паттернов и представление знаний

Когда применять Data Mining?

- не предназначен для проверки априорных предположений
- нужен, когда природа связей между переменными неизвестна («черный ящик»)
- учитывается и сравнивается большое число переменных
- для поиска закономерностей используются самые разные методы

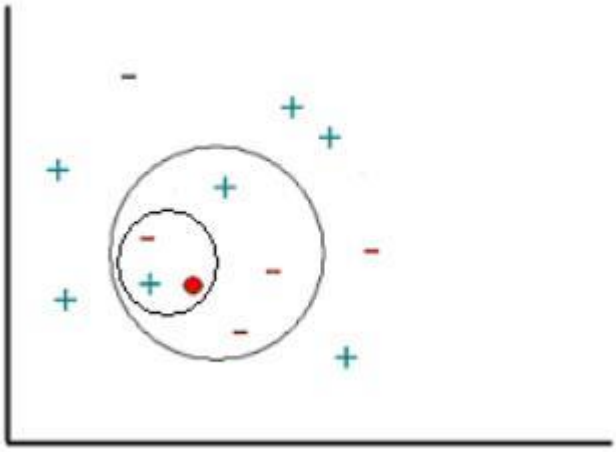
Типовые задачи, решаемые с помощью Data Mining

- Классификация
- Кластеризация
- Ассоциация
- Последовательность
- Прогнозирование
- Определение отклонений
- Анализ связей
- Визуализация

Специфика существующих алгоритмов

Алгоритм	Точность	Масшта-бируе-мость	Интер-претиру-емость	Пригод-ность	Трудо-ёмкость	Разносто-ронность	Быстрота	Популяр-ность
Линейная регрессия	3	4	3	4	3	3	4	2
Нейронные сети	4	2	2	2	3	2	1	2
Визуализация	4	1	4	4	5	2	1	3
Деревья решений	2	4	4	3	4	4	3	3
Полиномиальные нейронные сети	4	3	2	3	3	3	3	3
k-ближайшего соседа	2	1	3	3	3	2	4	2

Метод k-ближайших соседей – является одним из популярных нелинейных методов классификации. Основан на гипотезе компактности: схожие объекты чаще лежат в одном классе, чем в разных, если метрика расстояния между объектами введена достаточно корректно. Целевые переменные вычисляются на базе k «ближайших» к ним тренировочных экземпляров, определенных по формуле нахождения расстояния (например, по евклидовой метрике). Экземпляр относят к тому классу, который содержит большее число экземпляров-соседей.

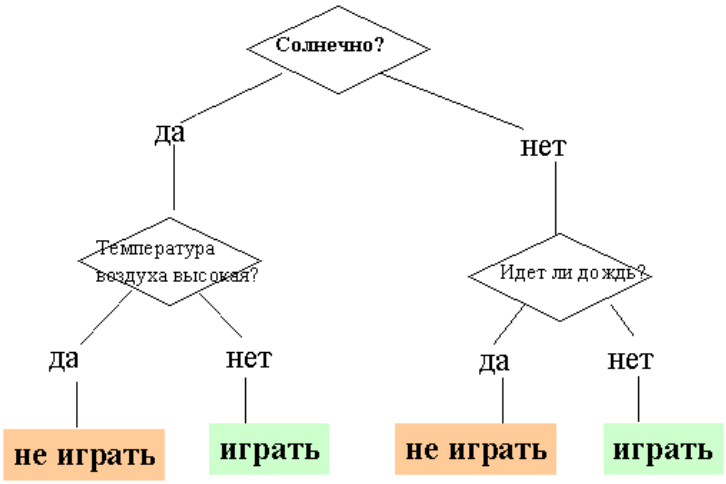


Искусственные нейронные сети (далее нейронные сети) часто применяют при решении задач классификации. В общем случае превосходит сложностью и требованиями к вычислительным ресурсам все остальные методы, но в определенных случаях дает отличные результаты. Основными методами этой группы являются: персептрон, включая его основные разновидности, и нейронные сети Кохонена, а именно сети векторного квантования, обучаемые с учителем.

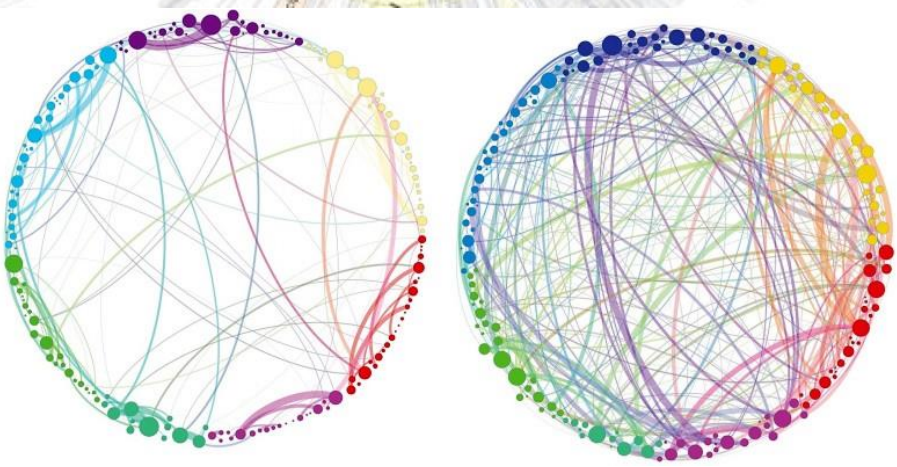
Нейронные сети также применяются для решения задачи кластеризации. Наиболее популярным методом из этого класса являются нейронные сети Кохонена – класс нейронных сетей, основным элементом которых является слой Кохонена, состоящий из линейных нейронов. Обычно выходные сигналы слоя Кохонена обрабатываются по принципу “Победитель получает все”, то есть, наибольший сигнал превращается в единичный, а остальные обращаются в ноль. Существует несколько разновидностей нейронных сетей Кохонена. Из них можно выделить два популярных метода для решения задачи кластеризации: сети векторного квантования, которые тесно связаны с статистическим алгоритмом k-means, и самоорганизующиеся карты Кохонена.

Логистическая регрессия – очень популярный линейный алгоритм, который распределяет наблюдения по категориям на базе количественных признаков. Считается простейшим алгоритмом машинного обучения для задач классификации и чаще всего используется для определения максимально возможной доли ошибок. Но в тоже время может дать достаточно хорошие результаты при определенных условиях. В результате предсказывает целевой класс или вероятность принадлежности к целевым классам.

Методы индукции правил, а именно: случайный лес и деревья решений, являются одними из популярных методов машинного обучения для классификации и регрессии. В процессе построения дерева решений, прогнозирующего целевую переменную, обучающую выборку рекурсивно разбивают на подмножества по результатам тестирования значения атрибута. Плюсом данного метода является то, что получаемое решение легко интерпретируется человеком. Случайный лес – “ансамбль” решающих деревьев, построенный по определенным правилам. В задаче классификации деревья решений “голосуют”. Как правило, дает более точные результаты, чем метод решающих деревьев. Случайный лес также является представителем композиционной группы, но, так как он объединяет в композицию только деревья решений, то его можно отнести к методам индукции правил.



Существуют инструменты визуализации данных в наборе, которые можно использовать для определения неявных зависимостей



При этом нельзя забывать, что Data Mining не может заменить аналитика.

- Data Mining имеет ряд существенных сложностей:
  - Высокие требования к квалификации пользователя
  - Сложность подготовки данных
  - Высокий процент бессмысленных результатов
  - Высокая стоимость
  - Наличие репрезентативных данных

Разница в подходах OLAP/DM

OLAP	Data Mining
Каковы средние показатели травматизма для курящих и некурящих?	Встречаются ли точные шаблоны в описании людей, подверженных травматизму?
Каково среднее соотношение существующих клиентов со счетами бывших клиентов?	Имеются ли характерные портреты клиентов, которые по всей видимости собираются отказаться от услуг связи?
Сколько в среднем совершают покупок по украденной и не украденной карточке	Существуют ли стереотипные схемы покупок для случая мошенничества с карточками?

Традиционная статистика, OLAP

Проверяют гипотезы, которые заранее сформулированы

Data Mining

Формируют новые гипотезы, обнаруживают неожиданные регулярности в данных, раскрывают hidden knowledge

Проблемы построения BI-систем

Необходимые данные недоступны

Низкое взаимодействие ИТ и пользователей

Отсутствие ясности у конечных пользователей

Данные для принятия решений поступают с задержкой

Несогласованность данных

Недостаточная подробность данных

Данные представляются в неудобных форматах

Медленная доставка данных

Данные невозможно выгрузить наружу

Низкое качество данных

Преждевременно агрегированные данные

Отвлечение на создание корпоративной модели данных

Использование всех доступных данных в системе