**Файловая система** – это система организации и хранения информации на жестком диске или других носителях, программные алгоритмы операционной системы для управления данной системой организации информации, и, наконец, на бытовом уровне это совокупность всех файлов и папок на диске.

Файловая система определяет:

- как хранятся файлы и каталоги на диске;

- какие сведения хранятся о файлах и каталогах;

- как можно узнать, какие участки диска свободны, а какие – нет;

- формат каталогов и другой служебной информации на диске.

Основные **функции** любой файловой системы нацелены на решение следующих задач:

* именование файлов;
* программный интерфейс работы с файлами для приложений;
* отображения логической модели файловой системы на физическую организацию хранилища данных;
* организация устойчивости файловой системы к сбоям питания, ошибкам аппаратных и программных средств;
* содержание параметров файла, необходимых для правильного его взаимодействия с другими объектами системы (ядро, приложения и пр.).

**БЗ**

**База данных** — организованная в соответствии с определёнными правилами и поддерживаемая в памяти компьютера совокупность [данных](https://dic.academic.ru/dic.nsf/ruwiki/71919), характеризующая актуальное состояние некоторой [предметной области](https://dic.academic.ru/dic.nsf/ruwiki/1327828) и используемая для удовлетворения информационных [потребностей](https://dic.academic.ru/dic.nsf/ruwiki/8874) пользователей. **База данных** – набор данных, распределенных по строкам и столбцам для удобного поиска, систематизации и редактирования.

**Документальные и фактографические базы данных.**

**Фактографическая ИС** - это массив фактов - конкретных значений данных об объектах реального мира. В системах **фактографического** типа в БД хранится информация об интересующих пользователя объектах предметной области в виде «фактов» (например, библиографические данные о сотрудниках, данные о выпуске продукции и т.п.). Информация в фактографической ИС хранится в четко структурированном виде, поэтому она способна давать однозначные ответы на поставленные вопросы, например: «Кто является победителем Чемпионата России по гимнастике в 1999 году?»,

В **документальных** БД единицей хранения является какой-либо документ (например, текст закона или статьи), и пользователю в ответ на его запрос выдается либо ссылка на документ, либо сам документ, в котором он может найти интересующую его информацию. Эту азу данных образует совокупность неструктурированных текстовых документов (статьи, книги, рефераты, тексты законов) и графических объектов, снабженная тем или иным формализованным аппаратом поиска. Цель системы, как правило, - выдать в ответ на запрос пользователя список документов или объектов, в какой-то мере удовлетворяющих сформулированным в запросе условиям. Например: выдать список всех статей, в которых встречается слово «Пушкин».

**Базы данных, хранящие данные в свободном формате.**

**Открытый формат файла** — общедоступная спецификация хранения цифровых данных, обычно разрабатываемая некоммерческой организацией по стандартизации, свободная от лицензионных ограничений при использовании. В частности, должна быть возможность включать поддержку открытых форматов.

**База данных свободного формата** – совокупность тематически связанных файлов в свободном формате. Отметим, что базы данных могут содержать файлы как в одном, так и в другом формате. Следовательно, внутримашинная информационная база представляет собой совокупность разнородных по формату баз данных.

Положительным качеством баз данных в свободном формате является возможность оперативного визуального обновления данных, предназначенных для некомпьютеризованного использования. Подобного рода информация применяется юристами, а также работниками налоговых служб, финансистами, бухгалтерами и др. В таких базах данных информация – это различные инструкции, законодательство, поправки к нему и пр.

[**Гипертекстовая база данных**](https://library_science.academic.ru/159) — Текстовая база данных, записи в которой содержат связи с другими записями, позволяющими компоновать комплексы записей на основе их логической связанности. При работе с гипертекстовой системой, пользователь имеет возможность просматривать документы (страницы текста) в том порядке, в котором ему больше нравится, а не последовательно, как это принято при чтении книг. Достигается это путем создания специального механизма связи различных страниц текста при помощи гипертекстовых ссылок

**Хранилище данных** – предметно-ориентированная информационная [база данных](https://ru.wikipedia.org/wiki/%D0%91%D0%B0%D0%B7%D0%B0_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85), специально разработанная и предназначенная для подготовки отчётов и бизнес-анализа с целью поддержки принятия решений в организации. Данные, поступающие в хранилище данных, как правило, доступны только для чтения.

**Витрина данных** представляет собой срез хранилища данных, представляющий собой комплекс тематически связанных баз данных, ориентированный, напрмер, на пользователей одной конкретной рабочей группы. Часто витрины называют киосками. Витрина данных является частью хранилища данных, специфицированной для использования конкретным подразделением или определенной группой пользователей.

**Технология Data mining** – интеллектуальная обработка данных с использованием методов машинного обучения, математической статистики и теории баз данных.

Data mining включает в себя обработку текстовых данных (text data mining) и графических и мультимедийных(web mining).

Устоявшегося перевода словосочетания «data mining» на русский язык нет, дословно это переводится как «добыча данных». Но чаще говорят «интеллектуальная обработка данных».

В настоящее время data mining является частью большего понятия – Big data, которое помимо обработки данных включает в себя их сбор и хранение.

**Технические характеристики**

Фундаментально data mining основывается на 3-х понятиях:

* Математическая статистика – является основой большинства технологий, используемых для data mining, например, кластерный анализ, регрессионный анализ, дискриминирующий анализ и пр.;
* Искусственный интеллект – воспроизведение нейронной сети мышления человека в цифровом виде;
* Машинное обучение – совокупность статистики и искусственного интеллекта, способствующая пониманию компьютерами данных, которые они обрабатывают для выбора наиболее подходящего метода или методов анализа.

В data mining используются следующие основные классы задач:·

* обнаружение отклонений – выявление данных, отличающихся по каким-либо параметрам из общей массы;
* обучение ассоциациям – поиск взаимосвязей между событиями;
* кластеризация – группирование наборов данных, без заранее известных шаблонов;
* классификация – обобщение известного шаблона для применения к новым данным;
* регрессия – поиск функции, отображающей набор данных с наименьшим отклонением;
* подведение итогов – отображение в сжатом виде исходной информации, включая предоставление отчетов и визуализацию.

**Кейсы применения.** Анализ данных по операциям с кредитными картами, анализ данных ЖКХ, программы карт лояльности в магазинах с учетом предпочтения покупателей, национальная безопасность (обнаружение вторжений), исследование генома человека – всего лишь небольшая часть возможных вариантов применения data mining.

***Text Mining*** - это набор технологий и методов, предназначенных для извлечения информации из текстов. Основная цель – дать возможность работать с большими объемами исходных данных за счет автоматизации процесса извлечения нужной информации.

**Задачи Text Mining**

Важная задача технологии Text Mining связана с извлечением из текста его характерных элементов или свойств, которые могут использоваться как метаданные документа, ключевых слов, аннотаций.

Другая важная задача состоит в отнесении документа к некоторым категориям из заданной схемы их систематизации. Text Mining также обеспечивает новый уровень семантического поиска документов.

Возможности Text Mining могут применяться при управлении знаниями для выявления шаблонов в тексте, для автоматического «выталкивания» или размещения информации по интересующим пользователей профилям, создавать обзоры документов.

**Основные элементы Text Mining**

В соответствии с уже сформированной методологии к основным элементам Text Mining относятся:

• классификация (classification)

Задача классификации - это задачу распознавания, где по некоторой контрольной выборке система относит новый объект к той или другой категории.

В существующих сегодня системах классификация применяется, например, в избирательном распространении новостей подписчикам.

• кластеризация (clustering)

Кластеризация в Text Mining рассматривается как процесс выделения компактных подгрупп объектов с близкими свойствами. Система должна самостоятельно найти признаки и разделить объекты по подгруппам.

• построение семантических сетей,

Построение семантических сетей или анализ связей, которые определяют появление дескрипторов (ключевых фраз) в документе для обеспечения навигации.

• извлечение фактов, понятий (feature extraction),

• суммаризация (summarization),

• ответ на запросы (question answering),

• тематическое индексирование (thematic indexing),

• поиск по ключевым словам (keyword searching).

**Web Mining –** это использование методов интеллектуального анализа данных для автоматического обнаружения веб-документов и услуг, извлечения информации из веб-ресурсов и выявления общих закономерностей в Интернете

Web Mining решает следующие **задачи:**

* описание посетителей сайта (кластеризация, классификация);
* описание посетителей, которые совершают покупки в интернет-магазине (кластеризация, классификация);
* определение типичных сессий и навигационных путей пользователей сайта (поиск популярных наборов, ассоциативных правил);
* определение групп или сегментов посетителей (кластеризация);
* нахождение зависимостей при пользовании услугами сайта (поиск ассоциативных правил).

**Этапы** в Web Mining**:**

* *входной этап (англ. input stage)* — получение «сырых» данных из источников (логи серверов, тексты электронных документов);
* *этап предобработки (англ. preprocessing stage) —* данные представляются в форме, необходимой для успешного построения той или иной модели;
* *этап моделирования (англ. pattern discovery stage);*
* *этап анализа модели (англ. pattern analysis stage)* — интерпретация полученных результатов.

**Web Content Mining** (Извлечение веб-контента) — процесс извлечения знаний из контента документов или их описания, доступных в Интернете. Он основано на сочетании возможностей информационного поиска, машинного обучения и интеллектуального анализа данных.

Web Content Miningописывает автоматический поиск информационных ресурсов в Интернете и включает в себя добычу содержимого из веб-данных. По сути, Web Content Mining является аналогом метода интеллектуального анализа данных для реляционных баз данных, так как существует возможность найти похожие типы знаний из неструктурированных данных, находящихся в веб-документах. Веб-документ может содержать несколько типов данных. Некоторые из них частично структурированные, такие как HTML-документы, некоторые более структурированные, такие как в базах данных, но большинство информации хранится в неструктурированных текстовых данных.

**Web Structure Mining** (Извлечение веб-структур) –процесс обнаружения структурной информации в Интернете. Данное направление рассматривает взаимосвязи между веб- страницами.

Web Structure Mining пытается обнаружить модель, лежащую в основе ссылочной структуры в Интернете. Модель основана на топологии гиперссылки с или без описания ссылки и может быть использована для классификации Веб-страницы и полезна для получения информации. Ссылочная структура содержит важную информацию, и может помочь в фильтрации и ранжировании веб-страниц. В частности, ссылка со страницы А на страницу В может считаться рекомендацией страницы B автором А.

**Web Usage Mining** — это процесс извлечения полезной информации из пользовательских журналов доступа, журналов прокси-сервера, браузерных журналов, пользовательских сессионных данных. Говоря простым языком, Web Usage Mining — это процесс выяснения того, что пользователи делают в Интернете. Это направление основано на извлечении данных из логов веб-серверов.

Статистика фиксирует идентификационные данные веб-пользователей вместе с их поведением на сайте. В зависимости от вида использования данных, результатом работы Web Usage Mining будут являться: данные веб-сервера; данные серверных приложений; данные прикладного уровня.

**OLAP (Online Analytical Processing)** – (аналитическая обработка данных в реальном времени) представляет собой мощную технологию обработки и исследования данных. Системы, построенные на основе технологии OLAP, предоставляют практически безграничные возможности по составлению отчетов, выполнению сложных аналитических расчетов, построению прогнозов и сценариев, разработке множества вариантов планов. OLAP системы могут организовать данные в соответствии с некоторым набором критериев. При этом не обязательно, чтобы критерии имели четкие характеристики.

Свое применение OLAP системы нашли во многих вопросах: стратегическое планирование, прогнозирование развития, подготовка финансовой отчетности, анализ работы и пр.

**Структура** OLAP системы

* база данных. База данных является источником информации для работы OLAP системы. Вид базы данных зависит от вида OLAP системы
* OLAP сервер. Он обеспечивает управление многомерной структурой данных и взаимосвязь между базой данных и пользователями OLAP системы.
* Пользовательские приложения. Этот элемент структуры OLAP системы осуществляет управление запросами пользователей и формирует результаты обращения к базе данных (отчеты, графики, таблицы и пр.)
* Виды систем

**Существует три способа хранения и обработки данных:**

* локально. Данные размещаются на компьютерах пользователей. Обработка, анализ и управление данными выполняется на локальных рабочих местах.
* реляционные базы данных. Данные хранятся на сервере этих систем в виде реляционных баз данных или хранилищ данных.
* многомерные базы данных. Данные организованы в виде специального хранилища данных на выделенном сервере. Все операции с данными осуществляются на этом сервере, который преобразует исходные данные в многомерные структуры. Такие структуры называют OLAP кубом.

**Виды OLAP систем**

**1. ROLAP** (Relational OLAP – реляционные системы).

**2. MOLAP** (Multidimensional OLAP – многомерные). Этот вид OLAP систем относится к традиционным системам. Отличие традиционной OLAP системы, от других систем, заключается в предварительной подготовке и оптимизации данных.

MOLAP системы являются самыми эффективными при обработке данных, т.к. они позволяют легко реорганизовать и структурировать данные под запросы пользователей.

**3. HOLAP** (Hybrid OLAP – гибридные OLAP системы). Гибридные OLAP системы представляют собой объединение систем ROLAP и MOLAP . В гибридных системах постарались объединить преимущества двух систем: использование многомерных баз данных и управление реляционными базами данных. HOLAP системы позволяют хранить большое количество данных в реляционных таблицах, а обрабатываемые данные размещаются в предварительно построенных многомерных OLAP кубах..

Д**ругие виды** OLAP систем (они являются больше маркетинговым ходом):

* WOLAP (Web OLAP)..
* DOLAP (Desktop OLAP).
* MobileOLAP.
* SOLAP (Spatial OLAP).

**Дата-центры (ЦОДы – центры обработки данных).**

Центр обработки данных, или дата-центр – это специализированное здание, в котором компании размещают свое серверное и сетевое оборудование. Именно дата-центры отвечают за физическую сохранность и бесперебойную работу оборудования.

**Функции** ЦОД:

* защита размещённого оборудования от воздействия окружающей среды;
* обеспечение оборудования качественным и бесперебойным электропитанием;
* отвод выделяемого тепла, вентиляция кондиционированным воздухом;
* управление физическим доступом к оборудованию, его охрана..

Грубо говоря, **ЦОД** — это не центр обработки данных, а центр размещения компьютерного оборудования.

Размещать своё вычислительное оборудование в ЦОД’е могут клиенты, использующие его для своих внутрикорпоративных нужд, провайдеры облачных услуг и владельцы центра, которые в этом случае выступают в качестве облачных провайдеров. Последнее обстоятельство и привносит некоторую путаницу в термин «центр обработки данных».

Надо понимать, что обеспечение работы ЦОД’а и управление компьютерным оборудованием, находящимся в центре - задачи разные, требующие разных профессиональных навыков и квалификации. Поэтому далеко не все владельцы центров обработки данных стремятся к оказанию облачных услуг.