1. **Теоретические основы баз данных и анализа маршрутов**
   1. **Базы данных**

База данных (БД) — именованная совокупность данных, отражающая состояние объектов и их отношений в рассматриваемой предметной области, или иначе БД — это совокупность взаимосвязанных данных при такой минимальной избыточности, которая допускает их использование оптимальным образом для одного или нескольких приложений в определенной предметной области. БД состоит из множества связанных файлов.

Система управления базами данных (СУБД) — совокупность языковых и программных средств, предназначенных для создания, ведения и совместного использования БД многими пользователями.

По степени универсальности различаются два класса СУБД — системы общего назначения и специализированные системы.

СУБД общего назначения не ориентированы на какую-либо конкретную предметную область или на информационные потребности конкретной группы пользователей. Каждая система такого рода реализуется как программный продукт, способный функционировать на некоторой модели ЭВМ в определенной операционной обстановке. СУБД общего назначения обладает средствами настройки на работу с конкретной БД в условиях конкретного применения.

В некоторых ситуациях СУБД общего назначения не позволяют добиться требуемых проектных и эксплуатационных характеристик (производительность, занимаемый объем памяти и прочее). Тем не менее создание специализированных СУБД весьма трудоемкий процесс и для того, чтобы его реализовать, нужны очень веские основания.

**1.1.1 Функции СУБД**

Управление данными во внешней памяти Данная функция предоставляет пользователям возможности выполнения самых основных операций, которые осуществляются с данными, — это сохранение, извлечение и обновление информации. Она включает в себя обеспечение необходимых структур внешней памяти как для хранения данных, непосредственно входящих в БД, так и для служебных целей, например для ускорения доступа к данным.

Управление транзакциями

Транзакция — это последовательность операций над БД, рассматриваемых СУБД как единое целое. Транзакция представляет собой набор действий, выполняемых с целью доступа или изменения содержимого базы данных. Примерами простых транзакций может служить добавление, обновление или удаление в базе данных сведений о некоем объекте. Сложная же транзакция образуется в том случае, когда в базу данных требуется внести сразу несколько изменений. Инициализация транзакции может быть вызвана отдельным пользователем или прикладной программой.

Восстановление базы данных

Одним из основных требований к СУБД является надежность хранения данных во внешней памяти. Под надежностью хранения понимается то, что СУБД должна быть в состоянии восстановить последнее согласованное состояние БД после любого аппаратного или программного сбоя. Обычно рассматриваются два возможных вида аппаратных сбоев: 

- мягкие сбои, которые можно трактовать как внезапную остановку работы компьютера (например, аварийное выключение питания); 

- жесткие сбои, характеризуемые потерей информации на носителях внешней памяти.

Поддержание надежности хранения данных в БД требует избыточности хранения данных, причем та часть данных, которая используется для восстановления, должна храниться особо надежно. Наиболее распространенным методом поддержания такой избыточной информации является ведение журнала изменений БД.

Поддержка языков

БД Для работы с базами данных используются специальные языки, называемые языками баз данных.

В современных СУБД обычно поддерживается единый интегрированный язык, содержащий все необходимые средства для работы с БД, начиная от ее создания, и обеспечивающий базовый пользовательский интерфейс с базами данных. Стандартным языком наиболее распространенных в настоящее время реляционных СУБД является язык SQL (Structured Query Language — язык структурированных запросов). Язык SQL позволяет определять схему реляционной БД и манипулировать данными.

Словарь данных

Одной из основополагающих идей рассмотренной выше трехуровневой архитектуры является наличие интегрированного системного каталога с данными о схемах, пользователях, приложениях и т. д. Системный каталог, который еще называют словарем данных, является, таким образом, хранилищем информации, описывающей данные в базе данных. Предполагается, что каталог доступен как пользователям, так и функциям СУБД. Обычно в словаре данных: содержится следующая информация:

- имена, типы и размеры элементов данных;

- имена связей;

- накладываемые на данные ограничения поддержки целостности;

- имена пользователей, которым предоставлено право доступа к данным;

- внешняя, концептуальная и внутренняя схемы и отображения между ними;

- статистические данные, например частота транзакций и счетчики обращений к объектам базы данных.

Управление параллельным доступом

Одна из основных целей создания и использования СУБД заключается в том, чтобы множество пользователей могло осуществлять параллельный доступ к совместно обрабатываемым данным. Параллельный доступ сравнительно просто организовать, если все пользователи выполняют только чтение данных, поскольку в этом случае они не могут помешать друг другу. Однако когда два или больше пользователей одновременно получают доступ к базе данных, конфликт с нежелательными последствиями легко может возникнуть, например, если хотя бы один из них попытается обновить данные.

СУБД должна гарантировать, что при одновременном доступе к базе данных многих пользователей подобных конфликтов не произойдет.

Управление буферами оперативной памяти

СУБД обычно работают с БД значительного размера. Понятно, что если при обращении к любому элементу данных будет производиться обмен с внешней памятью, то вся система будет работать со скоростью устройства внешней памяти. Практически единственным способом реального увеличения этой скорости является буферизация данных в оперативной памяти. В развитых СУБД поддерживается собственный набор буферов оперативной памяти с собственной дисциплиной замены буферов.

Контроль доступа к данным

СУБД должна иметь механизм, гарантирующий возможность доступа к базе данных только санкционированных пользователей и защищающий ее от любого несанкционированного доступа. В современных СУБД поддерживается один из двух широко распространенных подходов к вопросу обеспечения безопасности данных: избирательный подход или обязательный подход. В большинстве современных систем предусматривается избирательный подход, при котором некий пользователь обладает различными правами при работе с разными объектами. Значительно реже применяется альтернативный, обязательный подход, где каждому объекту данных присваивается некоторый классификационный уровень, а каждый пользователь обладает некоторым уровнем допуска.

Поддержка целостности данных

Термин целостность используется для описания корректности и непротиворечивости хранимых в БД данных. Реализация поддержки целостности данных предполагает, что СУБД должна содержать сведения о тех правилах, которые нельзя нарушать при работе с данными, и обладать инструментами контроля за тем, чтобы данные и их изменения соответствовали заданным правилам.

1.1.2 Языки баз данных

В СУБД поддерживается несколько специализированных по своим функциям подъязыков. Их можно разбить на две категории: 

язык определения данных БД — ЯОД {DDL — Data Definition Language);

язык манипулирования данными— ЯМД (DML — Data Manipulation, Language).

Язык определения данных

Язык определения данных — описательный язык, с помощью которого описывается предметная область: именуются объекты, определяются их свойства и связи между объектами. Он используется главным образом для определения логической структуры БД.

Схема базы данных, выраженная в терминах специального языка определения данных, состоит из набора определений.

Язык ЯОД используется как для определения новой схемы, так и для модификации уже существующей. Результатом компиляции ЯОД — операторов является набор таблиц, хранимый в системном каталоге, в котором содержатся метаданные — т. е. данные, которые включают определения записей, элементов данных, а также другие объекты, представляющие интерес для пользователей или необходимые для работы СУБД. Перед доступом к реальным данным СУБД обычно обращается к системному каталогу.

Языки манипулирования данными

Язык манипулирования данными содержит набор операторов манипулирования данными, т. е. операторов, позволяющих заносить данные в БД, удалять, модифицировать или выбирать существующие данные.

Множество операций над данными можно классифицировать следующим образом:

1. операции селекции;

2. действия над данными:

включение — ввод экземпляра записи в БД с установкой его связей; 

удаление — исключение экземпляра записи из БД с установкой новых связей;

модификация — изменение содержимого экземпляра записи и коррекция связей при необходимости.

Языки манипулирования данными делятся на два типа. Это разделение обусловлено коренным различием в подходах к работе с данными, а следовательно, различием в базовых конструкциях в работе с данными.

Первый тип — это процедурный ЯМД.

Второй тип — это декларативный (непроцедурный) ЯМД.

К процедурным языкам манипулирования данными относятся и языки, поддерживающие операции реляционной алгебры, которую основоположник теории реляционных баз данных Э. Ф. Кодд ввел для управления реляционной базой данных. Реляционная алгебра — это процедурный язык обработки реляционных таблиц, где в качестве операндов выступают таблицы в целом.

Декларативные языки предоставляют пользователю средства, позволяющие указать лишь то, какие данные требуются. Решение вопроса о том, как их следует извлекать, берет на себя процессор данного языка, работающий с целыми наборами записей.

Реляционные СУБД обычно включают поддержку непроцедурных языков манипулирования данными — чаще всего это бывает язык структурированных запросов SQL или язык запросов по образцу QBE.

В настоящее время нормой является поддержка декларативного языка SQL, в основе которого лежит реляционное исчисление, также введенное Э Коддом. Этот язык стал стандартом для языков реляционных баз данных, что позволяет использовать один и тот же синтаксис и структуру команд при переходе от одной СУБД к другой.

Следует отметить, что язык SQL имеет сразу два компонента: язык DDL (ЯОД) для описания структуры базы данных, и язык DML (ЯМД) для выборки и обновления данных.

Другим широко используемым языком обработки данных является язык QBE, который заслужил репутацию одного из самых простых способов извлечения информации из базы данных. Особенно это ценно для пользователей, не являющихся профессионалами в этой области Язык предоставляет графические средства создания запросов на выборку данных с использованием шаблонов Ответ на запрос также представляет собой графическую информацию.

Часть непроцедурного языка ЯМД, которая отвечает за извлечение данных, называется языком запросов Язык запросов можно определить как высокоуровневый узкоспециализированный язык, предназначенный для удовлетворения различных требований по выборке информации из базы данных.

**1.2 Теоретические основы теории графов**

Графом G называется совокупность из некоторого (обычно конечного) множества V , элементы которого называются вершинами, и некоторого выделенного подмножества E множества V 2 пар элементов множества V (называемых ребрами). Обычно подразумевается, что пары вершин неупорядочены (граф неориентированный) и элементы в каждой паре различны (нет петель). Если рассматриваются упорядоченные пары, граф называется ориентированным (или орграфом).

1.2.1 Основные алгоритмы теори графов.

Поиск в глубину

Поиск в глубину (англ. Depth-first search, DFS) — один из методов обхода графа. Стратегия поиска в глубину, как и следует из названия, состоит в том, чтобы идти «вглубь» графа, насколько это возможно. Алгоритм поиска описывается рекурсивно: перебираем все исходящие из рассматриваемой вершины рёбра. Если ребро ведёт в вершину, которая не была рассмотрена ранее, то запускаем алгоритм от этой нерассмотренной вершины, а после возвращаемся и продолжаем перебирать рёбра. Возврат происходит в том случае, если в рассматриваемой вершине не осталось рёбер, которые ведут в нерассмотренную вершину. Если после завершения алгоритма не все вершины были рассмотрены, то необходимо запустить алгоритм от одной из нерассмотренных вершин.

Поиск в ширину

Поиск в ширину (обход по уровням) – один из алгоритмов обхода графа. Метод лежит в основе некоторых других алгоритмов близкой тематики. Поиск в ширину подразумевает поуровневое исследование графа: вначале посещается корень – произвольно выбранный узел, затем – все потомки данного узла, после этого посещаются потомки потомков и т.д. Вершины просматриваются в порядке возрастания их расстояния от корня.

Пусть задан граф G=(V, E) и корень s, с которого начинается обход. После посещения узла s, следующими за ним будут посещены смежные с s узлы (множество смежных с s узлов обозначим как q; очевидно, что q⊆V, то есть q – некоторое подмножество V). Далее, эта процедура повториться для вершин смежных с вершинами из множества q, за исключением вершины s, т. к. она уже была посещена. Так, продолжая обходить уровень за уровнем, алгоритм обойдет все доступные из s вершины множества V. Алгоритм прекращает свою работу после обхода всех вершин графа, либо в случае выполнения наличествующего условия.

Алгоритм Ли

Алгоритм работает на дискретном рабочем поле (ДРП), представляющем собой ограниченную замкнутой линией фигуру, не обязательно прямоугольную, разбитую на прямоугольные ячейки, в частном случае — квадратные. Множество всех ячеек ДРП разбивается на подмножества: «проходимые» (свободные), т. е при поиске пути их можно проходить, «непроходимые» (препятствия), путь через эту ячейку запрещён, стартовая ячейка (источник) и финишная (приемник). Назначение стартовой и финишной ячеек условно, достаточно — указание пары ячеек, между которыми нужно найти кратчайший путь.

Алгоритм предназначен для поиска кратчайшего пути от стартовой ячейки к конечной ячейке, если это возможно, либо, при отсутствии пути, выдать сообщение о непроходимости.

Работа алгоритма включает в себя три этапа: инициализацию, распространение волны и восстановление пути.

Во время инициализации строится образ множества ячеек обрабатываемого поля, каждой ячейке приписываются атрибуты проходимости/непроходимости, запоминаются стартовая и финишная ячейки.

Далее, от стартовой ячейки порождается шаг в соседнюю ячейку, при этом проверяется, проходима ли она, и не принадлежит ли ранее меченной в пути ячейке.

Соседние ячейки принято классифицировать двояко: в смысле окрестности Мура и окрестности фон Неймана, отличающийся тем, что в окрестности фон Неймана соседними ячейками считаются только 4 ячейки по вертикали и горизонтали, в окрестности Мура — все 8 ячеек, включая диагональные.

При выполнении условий проходимости и непринадлежности её к ранее помеченным в пути ячейкам, в атрибут ячейки записывается число, равное количеству шагов от стартовой ячейки, от стартовой ячейки на первом шаге это будет 1. Каждая ячейка, меченая числом шагов от стартовой ячейки становится стартовой и из неё порождаются очередные шаги в соседние ячейки. Очевидно, что при таком переборе будет найден путь от начальной ячейки к конечной, либо очередной шаг из любой порождённой в пути ячейки будет невозможен.

Восстановление кратчайшего пути происходит в обратном направлении: при выборе ячейки от финишной ячейки к стартовой на каждом шаге выбирается ячейка, имеющая атрибут расстояния от стартовой на единицу меньше текущей ячейки. Очевидно, что таким образом находится кратчайший путь между парой заданных ячеек[6]. Трасс с минимальной числовой длиной пути, как при поиске пути в окрестностях Мура, так и фон Неймана может существовать несколько. Выбор окончательного пути в приложениях диктуется другими соображениями, находящимися вне этого алгоритма. Например, при трассировке печатных плат — минимумом линейной длины проложенного проводника.

Алгоритм Дейкстры

Алгоритм Дейкстры (англ. Dijkstra’s algorithm) — алгоритм на графах, изобретённый нидерландским учёным Эдсгером Дейкстрой в 1959 году. Находит кратчайшие пути от одной из вершин графа до всех остальных. Алгоритм работает только для графов без рёбер отрицательного веса. Алгоритм широко применяется в программировании и технологиях, например, его используют протоколы маршрутизации OSPF и IS-IS.

Каждой вершине из V сопоставим метку — минимальное известное расстояние от этой вершины до a. Алгоритм работает пошагово — на каждом шаге он «посещает» одну вершину и пытается уменьшать метки. Работа алгоритма завершается, когда все вершины посещены.

Инициализация. Метка самой вершины a полагается равной 0, метки остальных вершин — бесконечности. Это отражает то, что расстояния от a до других вершин пока неизвестны. Все вершины графа помечаются как непосещённые.

Шаг алгоритма. Если все вершины посещены, алгоритм завершается. В противном случае, из ещё не посещённых вершин выбирается вершина u, имеющая минимальную метку. Мы рассматриваем всевозможные маршруты, в которых u является предпоследним пунктом. Вершины, в которые ведут рёбра из u, назовём соседями этой вершины. Для каждого соседа вершины u, кроме отмеченных как посещённые, рассмотрим новую длину пути, равную сумме значений текущей метки u и длины ребра, соединяющего u с этим соседом. Если полученное значение длины меньше значения метки соседа, заменим значение метки полученным значением длины. Рассмотрев всех соседей, пометим вершину u как посещённую и повторим шаг алгоритма.

**1.3 Пространственный анализ**

К средствам пространственного анализа относятся различные процедуры манипулирования пространственными и атрибутивными данными, выполняемые при обработке запросов пользователя. (Например, операции наложения графических объектов, средства анализа сетевых структур или выделения объектов по заданным признакам).

Для каждого ГИС-пакета характерен свой набор средств пространственного анализа, обеспечивающий решение специфических задач пользователя, в тоже время можно выделить ряд основных функций, свойственных практически каждому ГИС-пакету. Это, прежде всего, организация выбора и объединения объектов в соответствии с заданными условиями, реализация операций вычислительной геометрии, анализ наложений, построение буферных зон, сетевой анализ.

1.3.1 Основные функции пространственного анализа данных

Выбор объектов по запросу: самой простой формой запроса является получение характеристик объекта указанного курсором на экране и обратная операция, когда изображаются объекты с заданными атрибутами. Более сложные запросы позволяют выбирать объекты по нескольким признакам, например по признаку удаленности одних объектов от других, совпадающие объекты, но расположенные в разных слоях и т. д.

Для выбора данных в соответствии с определенными условиями используются SQL- запросы. Для выполнения запросов разной сложности реализованы возможности использования при составлении запросов математических и статистических функций, а также географических операторов, позволяющих выбирать объекты на основании их взаимного расположения в пространстве (например, находится ли анализируемый объект внутри другого объекта или пересекается с ним).

Обобщение данных может проводиться по равенству значений определенного атрибута, в частности для зонирования территории. Еще один способ группировки - объединение объектов одного тематического слоя в соответствии с их размещением внутри полигональных объектов других тематических слоев.

Геометрические функции: к ним относят расчеты геометрических характеристик объектов или их взаимного положения в пространстве, при этом используются формулы аналитической геометрии на плоскости и в пространстве. Так для площадных объектов вычисляются занимаемые ими площади или периметры границ, для линейных - длины, а также расстояния между объектами и т.д.

Оверлейные операции (топологическое наложение слоев) являются одними из самых распространенных и эффективных средств. В результате наложения двух тематических слоев образуется другой дополнительный слой в виде графической композиции исходных слоев. Учитывая, что анализируемые объекты могут относиться к разным типам (точка, линия, полигон), возможны разные формы анализа: точка на точку, точка на полигон и т.д. Наиболее часто анализируется совмещение полигонов.

Построение буферных зон. Одним из средств анализа близости объектов является построение буферных зон. Буферные зоны - это районы (полигоны), граница которых отстоит на заданном расстоянии от границы исходного объекта. Границы таких зон вычисляются на основе анализа соответствующих атрибутивных характеристик. При этом ширина буферной зоны может быть как постоянной, так и переменной. Например, буферная зона вокруг источника электромагнитного излучения, будет иметь форму круга, а зона загрязнения от дымовой трубы завода с учетом розы ветров будет иметь форму близкую к эллипсу.

Сетевой анализ позволяет пользователю проанализировать пространственные сети связных линейных объектов (дороги, линии электропередач и т. д.). Обычно сетевой анализ служит для задач определения ближайшего, наиболее выгодного пути, определения уровня нагрузки на сеть, определение адреса объекта или маршрута по заданному адресу и другие задачи.

1.3.2 Анализ пространственного распределения объектов

Анализ пространственного распределения объектов. Фактически во многих случаях необходимо знать не только объем пространства, занимаемый объектами, но и расположение объектов в пространстве, которое может характеризоваться количеством объектов в определенной области, например, распределение численности населения. Наиболее распространены методы анализа распределения точечных объектов. Мерой точечного распределения служит плотность. Она определяется как результат деления числа точек на значение площади территории, на которой они расположены. Кроме плотности распределения можно оценить форму распределения. Точечные распределения встречаются в одном из четырёх возможных вариантов: равномерном (если число точек в каждой малой подобласти такое же, как и в любой другой подобласти), регулярном (если точки, разделённые одинаковыми интервалами по всей области, расположены в узлах сетки), случайном, кластерном (если точки собраны в тесные группы).

Точечные распределения могут описываться не только количеством точек в пределах подобластей. Часто анализируются локальные отношения внутри пар точек. Вычисление этого статистического показателя включает определение среднего расстояния до ближайшей соседней точки среди всех возможных пар ближайших точек. Данный метод позволяет оценить меру разреженности точек в распределении.

Распределение линий также оценивается по плотности. Обычно вычисления выполняются для сравнения разных географических областей, например по густоте гидрографической сети. Линии могут также оцениваться по близости и возможным пересечениям. Другими важными характеристиками являются ориентация, направленность и связанность.

Анализ распределения полигонов подобен анализу распределения точек, однако при оценке плотности определяют не количество полигонов на единицу площади, а относительную долю площади, занимаемой полигоном.