

# 1. ВВЕДЕНИЕ В ТЕОРИЮ БАЗ ДАННЫХ

## 1.1. ОСНОВНЫЕ ПОНЯТИЯ

Исторически сложились два основных направления использования вычислительной техники, первое из которых связано с проведением сложных преобразований над относительно небольшими объемами данных с простой структурой. Здесь компьютеры позволили быстрее проводить расчеты по вычислительно сложным алгоритмам. Подобные задачи дали толчок к созданию первых ЭВМ, их актуальность не снижается и сейчас.

Другое направление связано с созданием информационных систем (ИС). В них необходимо не только обрабатывать, но и хранить большие объемы данных со сложной внутренней структурой, обеспечивать быстрый поиск нужной информации. Создание подобных систем стало возможным после появления надежных, емких и быстродействующих устройств энергонезависимой памяти: в первую очередь речь идет о накопителях на жестких магнитных дисках. Классическим примером систем подобного типа являются системы резервирования железнодорожных и авиационных билетов. Последовательность операций, выполняемых при каждом заказе, относительно проста, но для корректного функционирования всей системы необходимо хранить и постоянно актуализировать большие объемы данных, выполнять в них поиск и т. п.

*Автоматизированная информационная система* – это функционирующий на основе ЭВМ комплекс, обеспечивающий сбор, хранение, актуализацию и обработку информации в целях поддержки какого-либо вида деятельности, т. е. автоматизированная ИС разрабатывается для определенной предметной области.

*Предметная область* – часть реального мира, подлежащая изучению с целью организации управления и, в конечном счете, автоматизации. Создавая ИС, мы, в некотором смысле, создаем информационную модель, позволяющую описать значимые характеристики реальных объектов и их взаимосвязи.

Автоматизированная ИС может функционировать самостоятельно или служить компонентом более сложной системы, например, автоматизированной системы управления предприятием. По типу хранимой и обрабатываемой информации выделяют два больших класса автоматизированных информационных систем: документальные и фактографические.

*Документальные системы* служат для работы с текстами на естественном языке – статьями, научными отчетами, текстами законодательных актов и т. д. Наиболее распространенным видом документальных систем являются информационно-поисковые системы, предназначенные для накопления и поиска документов на естественном языке. Их иногда еще называют полнотекстовыми базами данных.

Документы, хранящиеся в подобных системах, составляют поисковый массив документов системы. Для каждого документа формируется поисковый образ – некое формальное описание документа в терминах языка системы, которое отражает его содержание. Например, поисковый образ может быть сформирован указанием набора ключевых слов. Запрос пользователя представляется в виде поискового образа запроса, который сопоставляется с поисковыми образами хранимых документов. Отобранные в результате документы называются релевантными запросу.

*Фактографические системы* составляют другой большой класс автоматизированных информационных систем. Они оперируют фактическими данными, представленными в виде специальным образом организованных совокупностей записей. Именно им и посвящена основная часть данного курса, так как именно в фактографических системах в полной мере используются методы и инструменты теории баз данных. Фактографические системы, создаваемые средствами технологии баз данных, часто называют банками данных (см. определение ниже).

Иногда в дополнение к выделенным двум классам вводят понятие *лексикографических* баз данных и информационных систем, относя к ним различного рода словари и классификаторы.

В отечественных нормативных документах в сфере разработки баз данных даются следующие определения.

*Банк данных* (БнД) [1, 2] – это система специальным образом организованных данных (баз данных), программных, технических, языковых, организационно-методических средств, предназначенных для обеспечения централизованного накопления и коллективного многоцелевого использования данных.

В англоязычной литературе понятие, по сути близкое к банку данных, обозначается термином *система баз данных* (англ. database system).

*База данных* (БД) – именованная совокупность данных, отражающая состояние объектов и их отношений в заданной предметной области.

Базу данных можно рассматривать как электронную картотеку, хранилище для некоторого набора занесенных в компьютер данных. Операции над базой данных:

- добавить новые данные в БД;
- изменить существующие данные;
- удалить данные из БД;
- найти данные в БД;
- и т. д.

Базы данных организуются на основе различных моделей данных. Пример фрагмента БД реляционного типа представлен в табл. 1.1. Данные в этом случае организуются в виде реляционных таблиц, строки таблиц называют записями, а столбцы – полями или атрибутами.

Таблица 1.1.

#### Фрагмент реляционной БД

StudID	ФИО	Group
123	Иванов И.И.	382
124	Петров П.П.	382

Принципиально важной особенностью БД является то, что они содержат дополнительную служебную информацию о своей структуре, иначе говоря, являются «самодокументируемыми».

## 1.2. КОМПОНЕНТЫ СИСТЕМЫ БАЗ ДАННЫХ

Рассмотрим упрощенную схему системы баз данных (рис. 1.1). Она включает следующие основные компоненты [3]: данные, аппаратное обеспечение, программное обеспечение, пользователи.

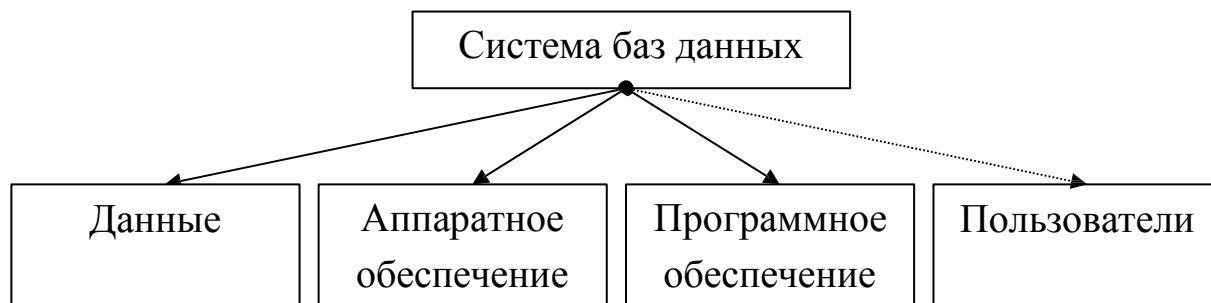


Рис. 1.1. Обобщенная схема системы баз данных

### Данные

Базы данных состоят из некоторого набора *постоянных данных*. Выделяют также *транзитные данные*, такие как промежуточные результаты, входные и выходные данные. *Входные данные* – информация, передаваемая системе (например, вводимая с клавиатуры). Такая информация может стать причиной изменений постоянных данных (она может стать частью постоянных данных), но не является частью БД как таковой. *Выходные данные* – сообщения и результаты, выдаваемые системой. Они, как правило, берутся из постоянных данных, но их нельзя рассматривать как часть БД.

Например, пусть в систему каждые 10 минут поступают данные о температуре воздуха, в базе сохраняется среднее значение за час, а запрос выводит среднесуточную температуру. В этом случае хранимые значения могут отличаться и от входных, и от выходных.

Кроме данных, описывающих предметную область, в БД обычно содержатся данные, описывающие элементы и структуры самой базы. Подобные описания относятся к разряду *метаинформации*, т. е. «информации об информации». Централизованное хранилище

метаинформации называется *словарем данных* или *репозиторием*. Именно наличие репозитория позволяет говорить о свойстве «самодокументированности» БД. В современных СУБД реляционного типа такое хранилище реализуется в виде системного каталога – набора служебных таблиц, куда заносится информация о структуре объектов (баз данных, таблиц, представлений и т.д.), пользователях, разрешениях и т.п.

По виду отношения «пользователь – данные» можно выделить два типа систем баз данных.

1. *Однопользовательская система* (англ. *single-user system*) – это система, в которой в одно и тоже время к базе данных может получить доступ только один пользователь.

2. *Многопользовательская система* (англ. *multi-user system*) – это система, в которой к базе данных могут получить доступ одновременно несколько пользователей. При этом для конечного пользователя необходимо обеспечить такие условия, чтобы результат его работы не зависел от того, работает он с данными в однопользовательском режиме или совместно с другими.

Данные в БД должны быть интегрированными и общими.

Когда говорят про *интегрированные данные*, подразумевают, что к данным, собранным из разных источников, предоставляется единый способ доступа. Например, система позволяет получить данные с кафедр университета об успеваемости студентов, из библиотеки – об использовании студентами литературы, и совместно их использовать для решения какой-то задачи.

*Общие данные* подразумевают возможность использования отдельных наборов данных из общей БД разными группами пользователей для решения своих специфических задач. Например, менеджер интернет-магазина может работать с данными о конкретном заказе, а руководитель – с итоговыми данными, характеризующими деятельность магазина за определенный период.

Эти два свойства представляют собой наиболее важное преимущество использования систем БД корпоративного уровня, а «интеграция» является преимуществом при использовании настольных (персональных) систем БД.

## **Аппаратное обеспечение**

В наиболее общем виде можно выделить две группы устройств, принципиально важных для систем баз данных. Во-первых, это устройства хранения данных. Во-вторых, устройства обработки данных. Для небольших систем и обработка, и хранение могут производиться на одном и том же компьютере. Крупная система баз данных может использовать различные типы систем хранения и множество серверов для обработки данных. Здесь возникает целый класс новых задач, связанных с разработкой и эксплуатацией распределенных систем.

## **Программное обеспечение**

Между физической базой данных и пользователями системы располагается уровень программного обеспечения, основной компонент которого – *система управления базами данных* (англ. database management system).

*Система управления базами данных* (СУБД) [2, 4] – совокупность языковых и программных средств, предназначенная для создания, ведения и совместного использования БД многими пользователями. Основная функция СУБД – предоставление пользователю БД возможности работать с ней, не вникая в детали на уровне аппаратного обеспечения.

Кроме СУБД система БД, как правило, включает еще ряд программных компонент – утилиты, генераторы отчетов, пользовательское прикладное программное обеспечение (ПО) и т. д.

## **Пользователи**

Пользователей системы БД можно разделить на три класса [3].

*Прикладные программисты* отвечают за написание прикладных программ, использующих базу данных. Разрабатываемые ими программы, обращаются с запросами к СУБД и получают результаты запросов. Выделяют программы пакетной обработки и оперативные приложения, функция которых – поддержка работы конечного пользователя, имеющего интерактивный доступ к системе.

*Конечные пользователи* работают с системой БД непосредственно с рабочей станции или терминала. Они могут воспользоваться разработанным для них прикладным ПО или встроенными средствами СУБД (графическими или с интерфейсом командной строки). Нужно понимать, что система БД создается для поддержания деятельности конечных пользователей.

*Администраторы данных и администраторы баз данных.* *Администратор данных* – человек, который несет ответственность за данные предприятия. Он принимает решения, какие данные необходимо вносить в БД, кому и к каким данным можно иметь доступ, и т. д. Иногда таких специалистов называют аналитиками. *Администратор базы данных* – технический специалист, который отвечает за реализацию решений администратора данных. На этапе разработки системы он занимается созданием баз данных, на этапе эксплуатации – настройкой, обслуживанием, резервным копированием и другими подобными задачами.

### **1.3. ЭТАПЫ РАЗВИТИЯ СУБД И ВЕДУЩИЕ ПРОИЗВОДИТЕЛИ**

До появления СУБД, вопросы хранения данных разработчики каждой программы решали самостоятельно, используя при этом функции операционной системы (ОС) или даже напрямую обращаясь к устройствам ввода-вывода. Но ОС предоставляет функции по работе с файлами, а вопросы организации хранения записей внутри файла, поиска данных, проверки ограничений для записи, средствами ОС не решить. Кроме того, при одновременном доступе нескольких пользователей к одним и тем же данным необходимы дополнительные механизмы, позволяющие централизованно управлять этим процессом. Эти и ряд других причин привели к созданию отдельного класса программного обеспечения – СУБД.

Первый этап развития СУБД связан с «большими» ЭВМ (мейнфреймами). Первая коммерческая СУБД называлась IMS (*от англ. Information Management System*, система управления информацией) и была выпущена корпорацией IBM в 1968 году для платформы IBM System/360. Этот этап характеризуется централизованным хранением

данных. СУБД должны были обеспечивать коллективный доступ к БД, а сами они работали на «больших» машинах под управлением сложных и достаточно развитых ОС.

На первом этапе исследователями были получены очень существенные результаты в области теории баз данных. В частности, это создание иерархической, сетевой и реляционной моделей данных. Реляционную модель предложил работавший в IBM математик Эдгар Франк Кодд (Edgar Frank Codd, 1923–2003; в 1981 получил премию Тьюринга). В 1970 году он опубликовал статью «A Relational Model of Data for Large Shared Data Banks», в которой описал основные идеи реляционного подхода. В дальнейшей работе над моделью принял участие и Кристофер Дейт (Christopher J. Date), автор классического учебника «Введение в системы баз данных» [3]. Реляционные СУБД на сегодняшний день являются наиболее распространенными.

Следующий этап развития СУБД связан с появлением персональных компьютеров. Их широкое распространение, ограниченные вычислительные возможности и, в среднем, более низкий (по сравнению с большими ЭВМ) уровень подготовки пользователей, привели к возникновению целого класса настольных СУБД. Изначально это были, в основном, однопользовательские системы, с достаточно ограниченными возможностями, но простым пользовательским интерфейсом и невысокими требованиями к аппаратуре. Многие из них не выдержали конкуренции и сейчас не поддерживаются. Оставшиеся в процессе развития стали приобретать черты многопользовательских СУБД, такие как механизмы совместного использования и защиты данных. В качестве примера популярных сейчас настольных СУБД можно назвать Microsoft Access и OpenOffice Base.

Параллельно существенные изменения происходили и с СУБД корпоративного уровня. Они были связаны с распространением компьютерных сетей, в результате чего доминирующей стала клиент-серверная технология, в том числе с поддержкой распределенной обработки данных.

Большое влияние на СУБД оказало и развитие сети Интернет. При динамическом формировании web-страниц в большинстве случаев

задействуются СУБД и обслуживаемые ими базы данных. Это привело к появлению ряда СУБД, чья популярность, в первую очередь, связана с их использованием при создании web-приложений. Наиболее яркий пример – реляционная СУБД MySQL.

С другой стороны, выяснилось, что реляционные СУБД и используемый для работы с ними язык запросов SQL подходят далеко не для всех задач. Появилась и активно развивается идеология NoSQL (*англ.* Not only SQL, не только SQL), объединяющая ряд подходов и проектов, связанных с созданием нереляционных БД.

Несколько слов об основных «игроках» на рынке баз данных. Наиболее именитый производитель серверных СУБД – это корпорация Oracle, выпустившая в 1979 году первую коммерческую реляционную СУБД Oracle v2, и с тех пор являющаяся ключевым производителем в области серверов баз данных.

Существенное место на рынке занимает корпорация IBM, выпускающая реляционную СУБД DB2 и иерархическую СУБД IMS. Приобретя в 2001 году подразделение корпорации Informix, IBM добавила в свою линейку продуктов одноименную СУБД.

Заметное место занимает корпорация Microsoft с ее серверным продуктом MS SQL Server и настольной СУБД Access, входящей в пакет Microsoft Office. Несмотря на то, что MS SQL Server выпускается только для ОС семейства Windows, популярность данной платформы, поддержка в средствах разработки Microsoft и широкие возможности самой СУБД, привели к её широкому распространению.

Основанная в 1984 году компания Sybase может быть также названа одним из пионеров в области разработки реляционных СУБД. В конце 1980-х – начале 1990-х Sybase вела разработку SQL Server в альянсе с Microsoft, но в дальнейшем продукты стали независимыми. На сегодняшний день в линейке продуктов Sybase есть реляционный сервер баз данных Adaptive Server Enterprise, встраиваемая реляционная СУБД SQL Anywhere и нереляционная СУБД с «поколоночным» хранением данных Sybase IQ, предназначенная для задач аналитической обработки данных и построения хранилищ данных. В 2010 году Sybase была приобретена компанией SAP AG, ведущим поставщиком программных решений для управления бизнесом.

Среди приверженцев свободно распространяемого программного обеспечения широкую популярность приобрела СУБД MySQL, изначально разрабатывавшаяся созданной в Швеции компанией MySQL AB. В настоящее время у MySQL лидирующие позиции в качестве СУБД, используемой в области web-разработки. В 2008 году компания MySQL AB была приобретена Sun Microsystems, а в 2010 году уже сама Sun приобрела Oracle. Сейчас выпускаются как коммерческие, так и бесплатно распространяемая версия MySQL (MySQL Community Edition). Кроме того, существуют разрабатываемые сообществом свободно распространяемые ответвления MySQL, например, это MariaDB.

Также необходимо отметить, что у многих коммерческих разработчиков есть бесплатно распространяемые версии СУБД, такие как Oracle Database Express Edition, IBM DB2 Express-C, Microsoft SQL Server Express Edition.

Если говорить о СУБД, основанных на объектной модели данных, то наиболее известным на сегодняшний день проектом в этой области является система Caché, разрабатываемая компанией InterSystems. Особенность данной СУБД заключается в том, что она реализует объектное представление данных, сохраняя в то же время возможность доступа к данным средствами языка SQL, как к реляционной БД.

## **БИБЛИОГРАФИЧЕСКИЙ СПИСОК**

1. Диго С.М. Базы данных. Проектирование и создание: Учебно-методический комплекс. – М.: Изд. центр ЕАОИ, 2008. – 171 с.
2. Диго С.М. Проектирование и использование баз данных: учебник для вузов по направлению и специальности "Информ. системы в экономике" / С. М. Диго.– М.: Финансы и статистика, 1995. – 207 с.
3. Дейт К. Дж. Введение в системы баз данных, 6-е издание: Пер. с англ. – К.; М.; СПб.: Издательский дом «Вильямс», 1999. – 848 с.
4. Карпова Т.С. Базы данных: модели, разработка, реализация. – СПб.: Питер, 2002 – 304 с.