Because learning changes everything.®

MIS

Databases

Issues with Traditional File Management Systems

Before the use of computers and database management systems (DBMS), manual file systems were used to maintain an organization's records and files.

Using this traditional file system:

- data and information were stored and processed using a traditional file system (paper, files, and documents)
- each file is independent of the other file, which leads to data redundancy, inconsistency, and file management issues

How Do DBMS Solve File Management Issues?

Data redundancy is the duplication of data, for example:

- if you are managing the data of a gym where a customer is enrolled in multiple workout classes, the same customer details will be stored twice, taking up storage and causing data redundancy
- data redundancy can lead to higher storage fees and inefficient access times

Data redundancy often leads to inconsistent data, to illustrate:

- if you are managing the data of a gym, let's say the customer needs to change their address
  - in a file management system, customer data is stored twice
- if all the address data is not changed in each record, data inconsistency will occur

How Do DBMS Solve File Management Issues? (cont.)

The use of a DBMS makes it is easier to secure data and information. DBMS allows for the creation of access constraints so that only authorized users are able to access the data.

Users are assigned a different set of access rules; this helps to protect data and users from identity theft, data leaks, and the misuse of data.

Database Management Systems (DBMS)

A **database management system (DBMS)** is a computer program that is used to create, process, and administer a database.

Due to the complexities surrounding the DBMS development process, most organizations do not develop their own DBMS.

Organizations use DBMSs created by software vendors including Oracle, Microsoft, and IBM.

Database Management Systems (DBMS) (cont.)

It is important to note the differences between a database and a DBMS.

Database

A database is a collection of tables, relationships, and metadata. DBMS helps to organize the data found in a database.

DBMS

A DBMS is a software program designed to organize and administer a database.

Data Integrity

Data integrity means the database is reliable, accurate, and aligned to the goals of the organization.

Data centralization is critical in increasing data integrity.

- when data is centralized, it means it is stored in only one place
- when multiple lists and data sources are maintained, information can become inconsistent leading to decreased data integrity

NicoElNino/Shutterstock

Relational Database Operations

Databases are designed to maintain data and information about various types of data objects including:

- objects (items in stock/inventory)
- events (transactions and item returns/exchanges)
- people (customers, employees, vendors)
- places (procurement centers and wholesalers)

Relational Databases

Relational databases organize data into tables based on structured data groupings.

Relational databases use links called relationships between tables.

- **tables** are used to hold information about the objects to be represented in the database
- information in tables is stored in rows called records or objects, and columns called fields. These relationships define how the data in the tables are related
- a common field that is included in both tables is used to create the relationship
- rows among multiple tables can be made related using foreign keys
- data can be accessed in many different ways without reorganizing the database tables themselves

Referential Integrity

Referential integrity is the accuracy and consistency of data within a table relationship in a database. In relational databases, two or more tables can be linked using a relationship.

The creation of relationships is achieved using:

- a primary key value (in the primary or parent table)
- foreign keys (in associated tables)

Referential Integrity (cont.)

Referential integrity requires that, whenever a foreign key value is used, it must reference a valid, existing primary key in the parent table.

For example, if you were to delete record 1556 in a primary table, you need to be sure that there is not a foreign key in any related table with a value of 1556.

You should only be able to delete a primary key if there are no associated records to that primary key.

If you delete the record you'll end up with an orphaned record.

Structured Query Language (SQL)

Structured Query Language (SQL), pronounced “ess-que-el,” is:

- used for human interface and communication with relational databases
- considered the standard language

SQL uses user-generated lines of code (statements) to answer questions against the database.

Most relational databases use SQL, but most also have proprietary extensions that allow for customized interactivity.

Data Normalization and Entity Relationship Diagrams (ERD)

Entities in Entity Relationship Diagrams (ERD) are the various business objects that make up the database and include:

- roles/people (employees and customers)
- tangible business objects (different types of products or services)
- intangible business objects (logs, system information)

Data normalization is a method of organizing various types of data in the database.

Normalization is an organized approach of breaking down/simplifying tables to eliminate data redundancy and undesirable data characteristics.

ERD are a method of structurally representing database design via the use of diagrams.

An ERD involves the use of different symbols and connectors that help to visualize two different types of information:

- The entities within the system
- The inter-relationships among these entities

NoSQL Databases

For many years, relational databases have been the most popular choice for businesses.

Due to increasing volumes of data, the increased use of web services, and the need for data storage, alternatives to relational databases are starting to emerge.

Instead of focusing on relational databases, some companies are turning to **non-relational databases (NoSQL)**.

- these databases are designed to manage large data sets across many platforms and have the ability to analyze structured and non-structured data
- they are also useful for creating queries from the data created from social media platforms, web apps, and other emerging forms of digital content
- a variety of NoSQL databases are available on the market today

NoSQL Databases (cont.)

Example

According to Amazon, DynamoDB can handle more than 10 trillion requests per day and can support peaks of more than 20 million requests per second.

Many of the world's fastest growing businesses, such as Lyft, Airbnb, and Redfin, as well as enterprises such as Samsung, Toyota, and Capital One, depend on the scale and performance of DynamoDB to support their mission-critical workloads.

Cloud Databases

A **Cloud database** is a type of database that is built and accessed via a Cloud platform.

A Cloud platform includes the hardware and operating environment of servers in an internet-based datacenter.

Key Features of Cloud Databases

- ability for enterprise users to host databases without having to buy and maintain dedicated hardware
- can be self-managed or maintained and managed by a provider
- support SQL and NoSQL databases
- accessed through the web or a vendor-provided API (Application Programming Interface)

Big Data

Big data encompasses all of the analysis tools and processes related to applying and managing large volumes of data.

Why Big Data?

Big Data was conceived out of the need of organizations to better understand trends, patterns, and preferences that emerge from the interaction with different systems and databases.

How does Big Data help?

Big data allows organizations to use analytics to help uncover a variety of predictive behaviors to help create new offerings.

The four common characteristics of big data include: Volume, Variety, Veracity, and Velocity

The For V's of Big Data

Volume: Scale of data

Enormous amounts of data are created every day. Most companies have over 120 terabytes of information stored (that's 120,000 gigabytes!).

Variety: Different forms of data

Data comes from many structured and unstructured sources. These sources include social media platforms, email, photos, videos, and point-of-sale interactions.

Veracity: Uncertainty of data

With all of the data being generated and stored it is important to ensure that data is meaningful and useful.

Velocity: Analysis of streaming data

The pace at which data is generated is mind-blowing. Ninety percent of the data on the Internet has been created since 2016.

Business Intelligence

Business Intelligence (BI) includes the technologies, computer applications, and procedures for the collection, analysis, and presentation of business information to help support decision making.

Key Ideas

- fundamentally, Business Intelligence systems are data-driven Decision Support Systems (DSS) that aid businesses to make better strategic decisions
- BI systems provide businesses a picture of historic, current, and future views of operations
- BI systems use information stored in data warehouses, data marts, in-memory computing, and other analytic platforms to create information output

Data Warehouse

A **data warehouse** is a repository of data and information that organizations analyze to make informed business and operational decisions.

Data and the analytics provided from the analysis of data allow organizations to create/maintain a competitive advantage.

Data Warehouse (cont.)

- Data flows into a data warehouse from a variety of transactional systems (point-of-sale, online transactions, etc.), databases, and other data-generating sources.
- Information flows into a data warehouse at regular intervals and is stored for later processing.
- A variety of people within an organization have access to the data warehouse, including data scientists, key decision-makers (KDMs), and data specialists.
- Data is analyzed using business intelligence (BI) tools, Structured Query Language (SQL) clients, and a variety of analytics applications designed to interpret the data.
- The output created from data warehouses includes reports, dashboards, and queries.

Data Mart

Whereas data warehouses are considered multi-purpose data and information storage facilities, a **data mart** is a subsection of a data warehouse that is designed and built specifically for individual departments or business functions.

Dependent Data Mart

- constructed from existing data warehouses and utilize a top-down approach where organizational data is stored in a centralized location, then specific data is extracted when analysis is needed

Independent Data Mart

- a stand-alone system that is created separate from a data warehouse and focuses on specific organizational functions

Hybrid Data Mart

- assimilates data from a data warehouse as well as other data collection systems.

Data Warehouses and Data Marts in Business

Data Warehouses

Data warehouses help to create a decision support system (DSS) environment which allows businesses to gauge the performance of an enterprise over measurable periods of time.

Use

Data warehouses contain large amounts of historical data (data is stored in a series of snapshots) that represent data points at a specific time. This allows organizations the ability to compare different time periods to make more informed business decisions.

- one of the advantages of data warehouses is their ability to provide access and analysis of information from a variety of subject areas

Data Warehouses and Data Marts in Business (cont)

Data Marts

Data marts are designed to collect and measure data from specific operational areas of a business and are used by individual departments or groups.

Use

Data marts are used to track inventories, purchase transactions, and the supply chain.

- Data marts assist with the analysis of what data a user needs rather than focusing on existing data

In-Memory Computing

In-memory computing is the use of middleware software to assist in the storage of data in random-access memory (RAM), across a group of different computers.

- Middleware is software that provides processing capabilities outside what is offered by the user's operating system
- RAM storage and parallel processing are two of the main components of in-memory computing.

In-Memory Computing (cont.)

Function: In-Memory Computing

In-memory is used by businesses to help unite transactional and analytical processing to provide real-time insights and analytics.

- this creates an environment that increases the amount and speed at which data can be ingested and analyzed

Analytic Platforms

Analytics platforms are designed to assist large data-driven companies in the analysis and interpretation of organizational data.

Purpose

Analytics platforms provide information about a variety of business and operational areas including:

- Customer analytics
- Sales and marketing analytics
- Social media analytics
- Cybersecurity
- Plant and facilities data

Analytic Platforms (cont.)

Example

One of the popular analytics platforms is IBM's Integrated Analytics System. According to IBM, "The IBM Integrated Analytics System":

- is a unified hybrid data management analytics solution providing massively parallel processing (MPP)
- comprises a high-performance hardware platform, optimized database query engine software and networking capabilities that work together to support various data analysis and business-reporting capabilities

Online Analytical Processing (OLAP)

Online Analytical Processing (OLAP) is included in many Business Intelligence (BI) software applications and is used for a variety of data discovery activities including report creation and analysis, analytical calculations, forecasting, budgeting, planning, and what-if predicative analysis.

OLAP software allows users to perform multidimensional analysis of a wide range of business data, complex calculations, and trend analysis, as well as data modeling.

Online Analytical Processing (OLAP) (cont.)

OLAP is used to assist businesses in a wide range of areas including:

- performance management
- financial reporting
- simulation models
- data warehouse reporting

Newly developed OLAP systems maintain a constant connection with back-end systems and allow for the delivery of reports and analytics in Microsoft Excel and other front-end tools used to collect data.

Data Mining

Data mining (also referred to as Knowledge Discovery in Data—KDD) is the searching of large data stores and sets to uncover patterns and trends that cannot be executed using OLAP or simple analysis techniques.

The four key properties of data mining include:

1. automatic discovery of patterns
2. prediction of likely outcomes
3. creation of actionable information
4. focus on large data sets and databases

Data Mining (cont.)

The data mining process includes:

- problem definition
- data gathering and preparation
- model building and evaluation
- knowledge deployment

Data Mining Information

Data mining is a means of analyzing data that can help organizations find patterns and relationships within data sets.

Key Ideas about Data Mining

- While data mining is a powerful tool, it does not replace the need to have an intimate knowledge of the organization, the data that is produced, and analytical methods employed to turn data into information
- Data mining assists businesses in uncovering information that may be hidden in data sets but does offer an organization why this information may be valuable
- Predictive information and relationships that are produced from data mining are not causal relationships
- Data mining yields probabilities, not exact answers

Data Mining Information (cont.)

Example

Data mining might determine that females with incomes between $75,000 and $100,000 who subscribe to certain female-targeted magazines may be more likely to purchase various products.

It is important that analysts not assume that the population identified through data mining buys the product because they belong to the identified population.

Web Mining

Web mining uses the principles of data mining to uncover and extract information from web sites, social media sites, e-commerce platforms, and web services.

Commonly Used Techniques to Gain Information

Technique	Description
Web Content Mining (WCM)	Includes extraction of information from web pages/documents, including text, images, videos, and interactives.
Web Structure Mining (WSM)	Includes analysis of hyperlinks, nodes, and related web pages.
Web Usage Mining (WUM)	Also called log mining, includes analysis of web access logs or the when, how, and frequency of web site access.

Web Mining Improves Web Experiences

The use of web mining by organizations can lead to improved web site visibility, usability, and accessibility.

- **Site visibility** includes how/when the site surfaces when queries are executed in search engines. Search Engine Optimization (SEO) can be enhanced through the information gained by web mining
- **Usability** refers to how easily web site users/visitors can interact with the site
- **Accessibility** includes the structure of web sites and pages to ensure device/platform access and scalability

Advantages of Accessing Databases via the Web

Access to internal databases via web technologies is increasingly important in today's competitive environment.

Advantages of using the web to access databases include:

1. The knowledge and ability of most users to operate web browsers and digital devices.
2. Front-end web interfaces require little change to existing database structures.
3. The increasing amount of global web connectivity and reach.

Information Policies

Database policies should be created and implemented by all organizations, regardless of the size of the organization.

Information policies specify:

- the rules used in database design (how data is structured)
- who has access to the data
- how the data is collected and maintained
- where information and data are distributed

Example

The information policy at a college/university would specify that only selected faculty and staff would have access to students' educational records. Only those people/areas that need access to this information would be able to granted access rights.

Data Administration

Data administration is responsible for the policies and procedures that are used to manage an organization's data.

Common data administration tasks include:

- development of information policies
- data planning
- database design
- security
- how internal-users and end-users use data

Data administration also includes the development of data modes and data dictionaries.

Data Governance

Data governance (DG) is an important component of a data management plan. Data governance includes the personnel, processes, and technology needed to oversee and secure an organization’s data and data assets.

Data governance policies help to ensure an organization's data is:

- Valid
- Understandable
- Complete
- Accessible

Data Governance (cont.)

Areas covered in digital governance include:

- Data architecture
- Data quality
- Data modeling
- Data warehousing and BI
- Data security

The key goals of data governance include:

- risk mitigation
- rules for data use
- compliance to requirements
- cost reduction
- high-quality data

Data Governance (cont. 2)

When creating a data governance plan, an organization should determine:

- Goals and objectives of the organization (Where/Who)
- Aspects of business governance should cover (What)
- Technical aspects and specifications (How)

Data Administrator

Database Administrators (DBA) is a technically specific role that is usually part of an organization's IT department.

Database administrators monitor and troubleshoot (when necessary) the database to ensure it is functional and available when needed.

Data Administrator (cont.)

Specific DBA tasks include:

- database security
- tuning
- backup
- creation of queries and reports that are used to assist business decisions

Data Quality Audit

Data that is of poor quality and validity has many risks. In order to ensure an organization's data are of the highest quality, a data quality audit can be used.

Method

A data quality audit uses statistical analysis to test data, variabilities, and outcomes against test data.

Purpose

Data quality audits help test the accuracy and completeness of an organization's data and provide a complete picture of the state of data and what improvements can be made to clean up data issues.

Data Quality Audit (cont.)

Key Benefits of Executing a Data Quality Audit

- reduced risk of data inconsistency
- reduced data storage costs
- data improvement recommendations

Use of Data Quality Audit

Conducting **data quality audits** helps to ensure data being used by business intelligence and other applications are of the highest quality.

Analyzing Data

Process Tasks	Description
Quality Assessment	Analyzes the quality of source data. Data sources including data warehouses and metadata are analyzed during this phase.
Data Design	Involves the creation of quality processes used to manage data.
Quality Transformation	Incorporates correction maps that are designed to correct issues that are present in source data.
Quality Monitoring	Uses an established process to examine data over a given amount of time to ensure data rules are being followed and that data is valid.

Data Scrubbing

Data scrubbing (data cleansing) includes the detection of errors in data sets and the removal/correction of these mistakes to ensure an organization's data are valid.

Examples of data errors data scrubbing can resolve include:

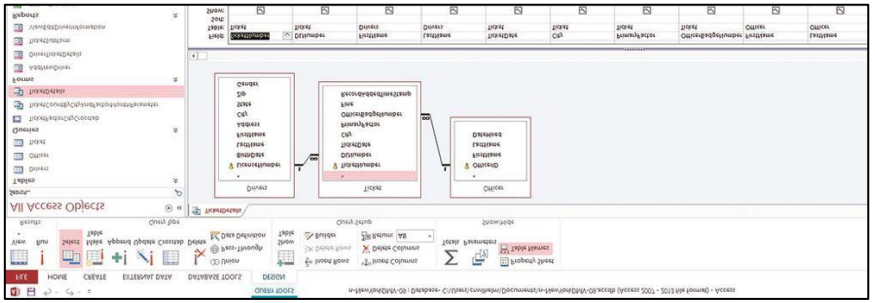
- elimination of duplicate database records
- correcting misspellings
- correcting incorrect names and address
- fixing syntax issues

Specialized data-scrubbing software is commonly used to conduct automatic surveys of data files and data sets.

What Is a Query?

A query is also known as a question.

- queries request information from a table (or combination of tables) in a database
- structured query language (SQL) is one of the most popular query languages used in creating queries
- queries allow you to find specific data quickly by filtering based on specific criteria, calculate/summarize data, and automate data management tasks



How to Run a Query in a Relational Database Program

Relational database management systems (RDBMS) have a variety of tools that can be used to execute/run queries.

Query types include:

- Crosstab: calculate the sum, average, or other aggregate functions, then group the results by two sets of values.
- Action: four types of action queries included in many RDBMS and include append, delete, update, and make-table queries.
- Parameters: prompt the user for values in order to run/execute the query.
- Structured Query Language (SQL): specific queries use specific SQL statements to execute the query.

Because learning changes everything.®

www.mheducation.com