**Databases and Data Acquisition :**

Databases categories:

* Relational
* Nonrelational

Relational Model:

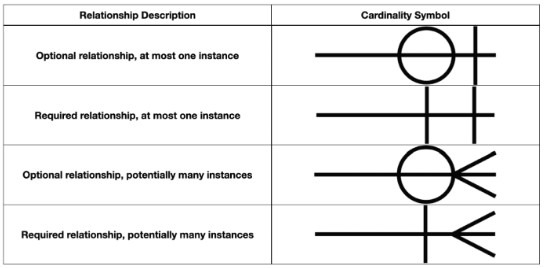
* 1969 – Edgar F. Codd from IBM developed relational Model for Database management.
* Builds concept of Tabular data.
* Entity contains data for a single subject.

**Cardinality** refers to the relationship between two entities, showing how many instances of one entity relate to instances in another entity.

ERD –

The first component of the terminator indicates whether the relationship between two entities is optional or required

The second component indicates whether an entity instance in the first table is associated with a single entity instance in the related table or if an association can exist with multiple entity instances.



A unary relationship is when an entity has a connection with itself.

A binary relationship connects two entities

A ternary relationship connects three entities.

**N:B:** Binary relationships are the most common and easy to explore, whereas unary and ternary are comparatively complex and rare.

A database administrator (DBA) is a highly trained person who understands how database software interacts with computer hardware.

* looks after how the database uses the underlying storage, memory, and processor resources assigned to the database
* looks for processes that are slowing the entire database down.

Referential Integrity

Relational Database Providers:

* oracle – released in 1979
* Microsoft SQL server
* open source community created offerings including MySQL, MariaDB, and PostgreSQL.
* Amazon Web Services (AWS) developed Aurora

Nonrelational Databases:

* Data validation happens in code, as opposed to being done in the database.
* Key-value - A key-value database is one of the simplest ways of storing data. Data is stored as a collection of keys and their corresponding values. A key must be globally unique across the entire database.
* Document - A document database is similar to a key-value database, with additional restrictions.
* Column-family = Column-family databases use an index to identify data in groups of related columns.
* Graph - Graph databases specialize in exploring relationships between pieces of data

**Database Use Cases:**

Databases tend to support two major categories of data processing:

* Online Transactional Processing (OLTP) - include booking a flight reservation, ordering something online, or executing a stock trade.
* and Online Analytical Processing (OLAP).

**Normalization** is a process for structuring a database in a way that minimizes duplication of data.

* One of the principles is that a given piece of data is stored once and only once. As a result, a normalized database is ideal for processing transactions.
* First normal form (1NF) is when every row in a table is unique and every column contains a unique value.
* 2NF applies an additional rule stating that all nonprimary key values must depend on the entire primary key.
* Third normal form (3NF) builds upon 2NF by adding a rule stating all columns must depend on only the primary key

**Online Analytical Processing**

* OLAP systems focus on the ability of organizations to analyze data.

**Schema concept**:

* The design of a database schema depends on the purpose it serves.
* Transactional systems require highly normalized databases, whereas a denormalized design is more appropriate for analytical systems.
* **Star schemas** are denormalized to improve read performance over large datasets.
* **A snowflake schema** is less denormalized than the star schema.
* One advantage of this approach is that it is easy to retrieve the current price while maintaining access to historical information.

**A data warehouse** is a database that aggregates data from many transactional systems for analytical purposes.

* A data warehouse facilitates analytics across the entire company.

Transactional data may come from systems that :

* power the human resources,
* sales,
* marketing,
* and product divisions.

Data Mart–

* a subnet of data warehouse.
* Data warehouses serve the entire organization, whereas data marts focus on the needs of a particular department within the organization.

Data Lake - stores raw data in its native format instead of conforming to a relational database structure.

**Dimensionality -** refers to the number of attributes a table has**.**

* he greater the number of attributes, the higher the dimensionality.
* One dimension you will frequently encounter is time. It is necessary to answer questions about when something happened or when something was true.

**Handling Dimensionality:**

* Another approach is to use an indicator flag for the current price.
* The indicator flag method keeps all pricing data in a single place
* It also simplifies the query structure to get the current price.

Data Acquisition Concepts

Integration - Data from transactional systems flow into data warehouses and data marts for analysis.

* **Approach** : extract, transform, and load (ETL)
* Extract – (First phase)
* extract data from the source system and place it in a staging area.
* The goal of the extract phase is to move data from a relational database into a flat file as quickly as possible.
* Transform – (Second phase)
* The goal is to reformat the data from its transactional structure to the data warehouse's analytical design.
* Load – (Last phase)
* The purpose of the load phase is to ensure data gets into the analytical system as quickly as possible.
* **Approach** : ELT(Extract, Load ,Transform), data is extracted from a source database and loaded directly into the data warehouse.

Difference:

* Component performing Data Transformation
* ETL -python
* ELT – SQL
* ELT has an advantage in the speed with which data moves from the operational to the analytical database.

**ELT Vendors:**

Data Collection:

* **APIs** - a structured method for computer systems to exchange information.
* APIs can be transactional, returning data as JSON objects.
* APIs can also facilitate bulk data extraction, returning CSV files.
* Web services
* **Web scrapping** - If data exists in a structured format, you can retrieve it programmatically. Programmatic retrieval of data from a website is known as web scraping.
* **Human-in-the-Loop** -There are times when the data you seek exists only in people's minds. (feelings)
* **Surveys** - One way to collect data directly from your customers is by conducting a survey.
* Survey tools
* **Observation** - Observation is the act of collecting primary source data, from either people or machines. Observational data can be qualitative or quantitative. Collecting qualitative observational data leads to unstructured data challenges
* Sampling

**Working with Data:**

DDL – Data Definition Language

* DDL lets you create, modify, and delete tables and other associated database objects.

DML - Data Manipulation Language

* capabilities of SQL to insert, modify, and retrieve information from databases.

\*\***While DDL manages the structure of a database, DML manages the data in the database.**

**Data Manipulation:**

**Four Actions to Occur:**

**CRUD (Create, Read, Update, Delete)**

1. Create new data.
2. Read existing data.
3. Update existing data.
4. Delete existing data.

DML Corresponding Verbs:

|  |  |  |
| --- | --- | --- |
| **Operation** | **SQL Keyword** | **Description** |
| **Create** | INSERT | Creates new data in an existing table |
| **Read** | SELECT | Retrieves data from an existing table |
| **Update** | UPDATE | Changes existing data in an existing table |
| **Delete** | DELETE | Removes existing data from an existing table |

Filtering with Logical Operators:

**AND operator**

SELECT Animal\_Name, Breed\_Name

FROM Animal

WHERE Animal\_Type = 'Dog'

AND Weight> 60

**OR operator**

SELECT Animal\_Name, Breed\_Name

FROM Animal

WHERE Animal\_Type = 'Dog'

OR Weight> 10

**Sorting :**

* **Oder by oldest**

SELECT Animal\_Name, Breed\_Name

FROM Animal

WHERE Animal\_Type = 'Dog'

AND Weight> 60

ORDER BY Date\_of\_Birth ASC

* **Order by youngest**

SELECT Animal\_Name, Breed\_Name

FROM Animal

WHERE Animal\_Type = 'Dog'

AND Weight> 60

ORDER BY Date\_of\_Birth DESC

SQL considerations:

* The keywords in SQL are case-insensitive. However, the case-sensitivity of column names and values depend on the database configuration

**IFF Function** 3 parameters:

1. Boolean Expression- returns true or false
2. True Value – returns true
3. False Value - returns true

**AGGREGATE** Functions :

* aggregate functions are an easy way to summarize data.
* Aggregate functions summarize a query's data and return a single value.

|  |  |
| --- | --- |
| **Function** | **Purpose** |
| **COUNT** | Returns the total number of rows of a query. |
| **MIN** | Returns the minimum value from the results of a query. Note that this works on both alphanumeric and numeric data types. |
| **MAX** | Returns the maximum value from the results of a query. Note that this works on both alphanumeric and numeric data types. |
| **AVG** | Returns the mathematic average of the results of a query. |
| **SUM** | Returns the sum of the results of a query. |
| **STDDEV** | Returns the sample standard deviation of the results of a query. |

**SYSTEM** Functions:

* Returns data about database environment

**Example:**

* whenever a person or automated process uses data from a database, they need to establish a database session.
* A database session begins when a person/program connects to a database. The session lasts until the person/program disconnects**.**
* a poorly written query can consume most of the resources available to the database. When that happens, a database administrator can identify and terminate the problematic session

**QUERY Optimization:**

* Parametrization - Effective use of parameterization reduces the number of times the database has to parse individual queries.
* Indexing - A database index works like the index in the back of a book. Instead of looking at each page in a book to find what you are looking for, you can find a specific page number in the index and then go to that page
* A database index can point to a single column or multiple columns.
* Data subnets and temporary tables
* Execution plan
* An execution plan shows the details of how a database runs a specific query.
* Execution plans are extremely helpful in troubleshooting query performance issues.
* They provide additional information about how a query is spending its time.