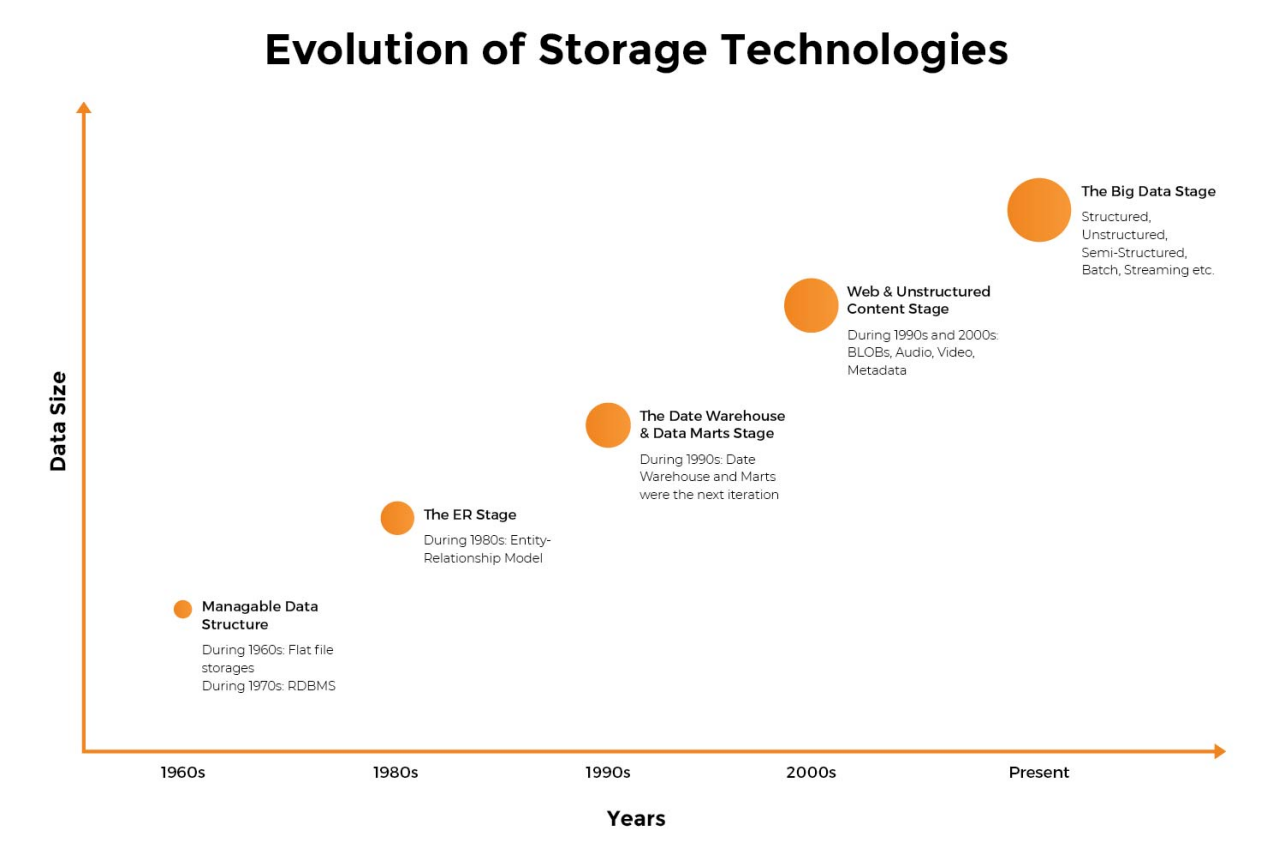
Relational vs Non-Relational

This lesson talks about the database landscape as it exists today in the industry.

Relational vs Non-Relational

There is a plethora of terms and jargon that can be confusing when first starting to read about data and its storage. We'll provide clarification on various terms below:



Data

Data (plural of datum) is defined as distinct pieces of information. Data can exist in several different forms: numbers, text, bytes, Instagram pictures, or YouTube videos. These represent various types of data that can be stored and transmitted electronically. Note that data is usually interpreted in a context, e.g., a data representing prose in the Hindi

language can't be interpreted as a picture and vice versa.

There are broadly three categories of data:

1. **Structured Data:** has some pre-defined organizational property about it that makes it easily searchable and analyzable. The data is backed by a model that dictates the size of each field of the data: its type, length, and any restrictions on what values it can take on. Data stored in SQL databases is structured. Structured data is usually formatted in a universally understandable and identifiable manner. In most instances, a schema formally specifies structured data. Whenever you are working in any variant of SQL, you are almost always dealing with structured data.

2. **Unstructured Data:** is characterized by a lack of organization and a data model that describes the structure of a single record or attributes of any individual fields within the record. Videos, audio, blogs, log files, social-media posts, etc, are all examples of unstructured data. It is data without any conceptual definition or type.

3. **Semi-structured Data:** is a cross between structured and unstructured data, and though there's no explicit data model or structure definition, one may be implied. Semi-structured data contains semantic tags, but does not conform to the structure associated with typical relational databases. Examples include JSON and XML data. It sits in between the spectrum of structured and unstructured data. Another example to consider is that of metadata related to videos and audios files, which themselves fall under unstructured data. The metadata such as creator, last accessed, permissions etc can be considered as semi-structured data.

   Semi-structured data contains certain parts that are structured, and others that are not. For example, X-rays and other large images consist largely of unstructured data comprising of millions of pixels. It is impossible to search and query these X-ray images in the same way that a large relational database can be searched, queried and analyzed. However even though the files themselves consist of only

pixels there exists a small section known as metadata within each file that consists of details such as author, creation timestamp, last accessed etc. This metadata allows for some form of analysis of unstructured data.

This brings us to the question: if all of unstructured data has associated metadata then what is the difference between semi-structured and unstructured data? In today's world there's hardly any truly unstructured data with no organization or metadata. In fact, the distinction between semi-structured and unstructured data is sort of a grey area and disputed. However, both are far away from the structured rigorously organized data living in relational databases. Often what is referred to as unstructured data such as videos, images, documents, social media postings, files etc is really semi-structured data. However, for simplicity data is usually divided and described as structured or unstructured.

In this course, we'll be exclusively dealing with structured data.

Database Management System (DBMS)

Usually, when we refer to databases such as MySQL and PostgreSQL, we are talking about a wholesome system, called the database management system. It's a software that allows the user to create, maintain, and delete multiple individual databases. It provides peripheral services and interfaces for the end-user to interact with the databases.

Database

A database is an organized and structured collection of data, usually stored and retrieved electronically. The structure and organization of data helps in efficient retrieval.

There are broadly two types of databases:

1. Relational or SQL Databases
2. Non-Relational or NoSQL Databases

In this course, our focus will be on relational databases.

Relational/SQL Databases

Relational databases consist of data stored as rows in tables. The columns of a table rigidly follow a defined schema that describes the type and size of the data that a table column can hold. You can think of a schema as a blueprint of what each record or row in the table should look like. This is why relational databases only handle structured data. Tables usually have a column as the key, which is used to uniquely identify each row in a table. A relationship between two tables is defined by a column or a set of columns that occur in both the tables.

Relational database management systems (RDBMS) are mature and widely adopted technology. Popular implementations include Oracle, DB2, Microsoft SQL Server, PostgreSQL, and MySQL.

Non-Relational/NoSQL Databases

The rise of Web 2.0 companies made NoSQL databases popular. As data sets handled by internet companies grew ever bigger in size, a new approach to designing databases came to the fore. The strict schema of a relational database was shunned in favor of a schema-less database. NoSQL databases come in different forms and address different use cases. The spectrum includes key-value stores (Redis, Amazon Dynamo DB), column stores (HBase, Cassandra), document stores (Mongo DB, Couchbase), graph databases (Neo4J), and search engines (Solr, ElasticSearch, Splunk). The primary distinction between these NoSQL databases and SQL databases is the absence of a rigid schema in the former. NoSQL databases, unstructured, and semi-structured data fall under the purview of Big Data.

Big Data

Big Data usually includes data sets with sizes beyond the capability of traditional software tools (e.g., SQL technologies) to capture, curate, manage, and process data within a tolerable elapsed time. The "big" in the term Big Data is a moving target that changes to a higher number as software and hardware capabilities enhance to process higher volumes of data.