Carnegie Mellon University

Intro to Database Systems (15-445/645)

Lecture #01

Relational Model & Algebra

FALL 2023 » Prof. Andy Pavlo • Prof. Jignesh Patel

SingleStore

COURSE LOGISTICS

Course Policies + Schedule: Course Web Page

Discussion + Announcements: Piazza

Homeworks + Projects: Gradescope

Final Grades: Canvas

Non-CMU students can complete all assignments using Gradescope (Code: **KK5DVJ**).

→ Do **not** post your solutions on Github.

→ Do **not** email instructors / TAs for help.

→ Discord Channel: https://discord.gg/YF7dMCg

→ Somebody needs to finish Andy's Wikipedia article.

COURSE LOGISTICS

Pavlo was born and raised in the streets of Baltimore, MD. After completing bachelor's and master's degrees from Rochester Institute of Technology and Brown University, he completed his Ph.D. from Brown University under Stan Zdonik and Mike Stonebraker.

Grades: Canvas

- Comment: "On the streets of..." Is not encyclopedic enough. Please use a NPOV voice when writing. And while I believe he is probably notable enough for an article- the sources here do not support that. we need signifcant coverage in independent secondary sources (IE- not somewhere he works OR a press release). Add these and do a bit of refining on the writing style and should be good to go! Nightenbelle (talk) 15:28, 9 July 2021 (UTC)

using

→ Do **not** post your solutions on Github.
→ Do **not** email instructors / TAs for help.
→ Discord Channel: https://discord.gg/YF7dMCg
→ Somebody needs to finish Andy's Wikipedia article.

LECTURE RULES

Do interrupt us for the following reasons:

→ We are speaking too fast.
→ You don't understand what we are talking about.
→ You have a database-related question.

Do **not** interrupt us for the following reasons:

→ Whether you can use the bathroom.
→ Questions about blockchains.

We will **not** answer questions about the lecture immediately after class.

TODAY'S AGENDA

Database Systems Background

Relational Model

Relational Algebra

Alternative Data Models

Databases


$7,400

DATABASE

Organized collection of inter-related data that models some aspect of the real-world.

Databases are the core component of most computer applications.

DATABASE EXAMPLE

Create a database that models a digital music store to keep track of artists and albums.

Things we need for our store:
→ Information about Artists
→ What Albums those Artists released

FLAT FILE STRAWMAN

Store our database as comma-separated value (CSV) files that we manage ourselves in our application code.

→ Use a separate file per entity.
→ The application must parse the files each time they want to read/update records.

Artist(name, year, country)

```
"Wu-Tang Clan",1992,"USA"
"Notorious BIG",1992,"USA"
"GZA",1990,"USA"
```

Album(name, artist, year)

```
"Enter the Wu-Tang","Wu-Tang Clan",1993
"St.Ides Mix Tape","Wu-Tang Clan",1994
"Liquid Swords","GZA",1990
```

FLAT FILE STRAWMAN

Example: Get the year that GZA went solo.

Artist(name, year, country)

```
"Wu-Tang Clan",1992,"USA"
"Notorious BIG",1992,"USA"
"GZA",1990,"USA"
```

```
for line in file.readlines():
  record = parse(line)
  if record[0] == "GZA":
    print(int(record[1]))
```

FLAT FILES: DATA INTEGRITY

How do we ensure that the artist is the same for each album entry?

What if somebody overwrites the album year with an invalid string?

What if there are multiple artists on an album?

What happens if we delete an artist that has albums?

FLAT FILES: IMPLEMENTATION

How do you find a particular record?

What if we now want to create a new application that uses the same database? What if that application is running on a different machine?

What if two threads try to write to the same file at the same time?

FLAT FILES: DURABILITY

What if the machine crashes while our program is updating a record?

What if we want to replicate the database on multiple machines for high availability?

DATABASE MANAGEMENT SYSTEM

A database management system (**DBMS**) is software that allows applications to store and analyze information in a database.

A general-purpose DBMS supports the definition, creation, querying, update, and administration of databases in accordance with some data model.

DATA MODELS

A data model is a collection of concepts for describing the data in a database.

A schema is a description of a particular collection of data, using a given data model.

DATA MODELS

Relational
Key/Value
Graph
Document / XML / Object
Wide-Column / Column-family
Array / Matrix / Vectors
Hierarchical
Network
Multi-Value

DATA MODELS

Relational ← Most DBMSs

Key/Value

Graph

Document / XML / Object

Wide-Column / Column-family

Array / Matrix / Vectors

Hierarchical

Network

Multi-Value

DATA MODELS

Relational
Key/Value
Graph
Document / XML / Object
Wide-Column / Column-family
Array / Matrix / Vectors
Hierarchical
Network
Multi-Value

← NoSQL

DATA MODELS

Relational
Key/Value
Graph
Document / XML / Object
Wide-Column / Column-family
Array / Matrix / Vectors ← Machine Learning
Hierarchical
Network
Multi-Value

DATA MODELS

Relational

Key/Value

Graph

Document / XML / Object

Wide-Column / Column-family

Array / Matrix / Vectors

Hierarchical

Network

Multi-Value

← Obsolete / Legacy / Rare

DATA MODELS

Relational ← This Course

Key/Value

Graph

Document / XML / Object

Wide-Column / Column-family

Array / Matrix / Vectors

Hierarchical

Network

Multi-Value

EARLY DBMSs

Early database applications were difficult to build and maintain on available DBMSs in the 1960s.

→ Examples: IDS, IMS, CODASYL

→ Computers were expensive, humans were cheap.

Tight coupling between logical and physical layers.

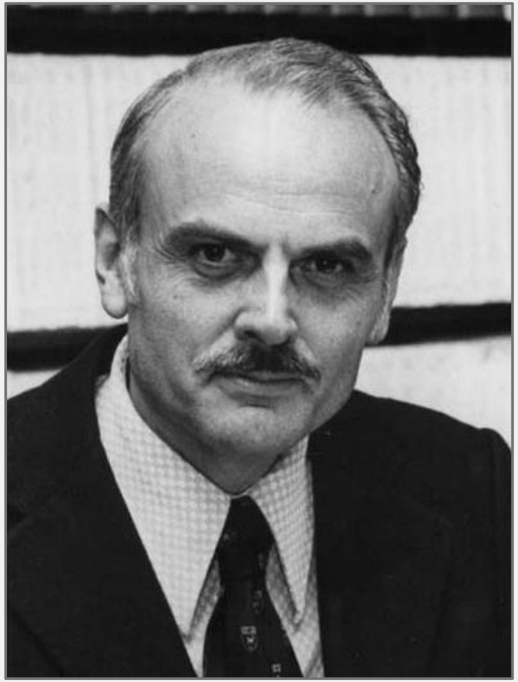
Programmers had to (roughly) know what queries the application would execute before they could deploy the database.

EARLY DBMSs

Ted Codd was a mathematician at IBM Research in the late 1960s.

Codd saw IBM's developers rewriting database programs every time the database's schema or layout changed.

Devised the relational model in 1969.



Edgar F. Codd

Y DB

T… n at
IB… .

C… writing
da… he
da… nged.

De… 1969.

DERIVABILITY, REDUNDANCY AND CONSISTENCY OF RELATIONS
STORED IN LARGE DATA BANKS

E. F. Codd
Research Division
San Jose, California

ABSTRACT: The large, integrated data banks of the future will contain many relations of various degrees in stored form. It will not be unusual for this set of stored relations to be redundant. Two types of redundancy are defined and discussed. One type may be employed to improve accessibility of certain kinds of information which happen to be in great demand. When either type of redundancy exists, those responsible for control of the data bank should know about it and have some means of detecting any "logical" inconsistencies in the total set of stored relations. Consistency checking might be helpful in tracking down unauthorized (and possibly fraudulent) changes in the data bank contents.

RJ 599(# 12343) August 19, 1969

Information Retrieval **P. BAXENDALE, Editor**

A Relational Model of Data for Large Shared Data Banks

E. F. Codd
IBM Research Laboratory, San Jose, California

Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation are changed. Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information.

Existing noninferential, formatted data systems provide users with tree-structured files or slightly more general network models of the data. In Section 1, inadequacies of these models are discussed. A model based on *n*-ary relations, a normal form for data base relations, and the concept of a universal data sublanguage are introduced. In Section 2, certain operations on relations (other than logical inference) are discussed and applied to the problems of redundancy and consistency in the user's model.

KEY WORDS AND PHRASES: data bank, data base, data structure, data organization, hierarchies of data, networks of data, relations, derivability, redundancy, consistency, composition, join, retrieval language, predicate calculus, security, data integrity
CR CATEGORIES: 3.70, 3.73, 3.75, 4.20, 4.22, 4.29

1. Relational Model and Normal Form

1.1. Introduction

This paper is concerned with the application of elementary relation theory to systems which provide shared access to large banks of formatted data. Except for a paper by Childs [1], the principal application of relations to data systems has been to deductive question-answering systems. Levein and Maron [2] provide numerous references to work in this area.

In contrast, the problems treated here are those of *data independence*—the independence of application programs and terminal activities from growth in data types and changes in data representation—and certain kinds of *data inconsistency* which are expected to become troublesome even in nondeductive systems.

The relational view (or model) of data described in Section 1 appears to be superior in several respects to the graph or network model [3, 4] presently in vogue for non-inferential systems. It provides a means of describing data with its natural structure only—that is, without superimposing any additional structure for machine representation purposes. Accordingly, it provides a basis for a high level data language which will yield maximal independence between programs on the one hand and machine representation and organization of data on the other.

A further advantage of the relational view is that it forms a sound basis for treating derivability, redundancy, and consistency of relations—these are discussed in Section 2. The network model, on the other hand, has spawned a number of confusions, not the least of which is mistaking the derivation of connections for the derivation of relations (see remarks in Section 2 on the "connection trap").

Finally, the relational view permits a clearer evaluation of the scope and logical limitations of present formatted data systems, and also the relative merits (from a logical standpoint) of competing representations of data within a single system. Examples of this clearer perspective are cited in various parts of this paper. Implementations of systems to support the relational model are not discussed.

1.2. Data Dependencies in Present Systems

The provision of data description tables in recently developed information systems represents a major advance toward the goal of data independence [5, 6, 7]. Such tables facilitate changing certain characteristics of the data representation stored in a data bank. However, the variety of data representation characteristics which can be changed *without logically impairing some application programs* is still quite limited. Further, the model of data with which users interact is still cluttered with representational properties, particularly in regard to the representation of collections of data (as opposed to individual items). Three of the principal kinds of data dependencies which still need to be removed are: ordering dependence, indexing dependence, and access path dependence. In some systems these dependencies are not clearly separable from one another.

1.2.1. *Ordering Dependence.* Elements of data in a data bank may be stored in a variety of ways, some involving no concern for ordering, some permitting each element to participate in one ordering only, others permitting each element to participate in several orderings. Let us consider those existing systems which either require or permit data elements to be stored in at least one total ordering which is closely associated with the hardware-determined ordering of addresses. For example, the records of a file concerning parts might be stored in ascending order by part serial number. Such systems normally permit application programs to assume that the order of presentation of records from such a file is identical to (or is a subordering of) the

RELATIONAL MODEL

The relational model defines a database abstraction based on relations to avoid maintenance overhead.

Key tenets:
→ Store database in simple data structures (relations).
→ Physical storage left up to the DBMS implementation.
→ Access data through high-level language, DBMS figures out best execution strategy.

RELATIONAL MODEL

Structure: The definition of the database's relations and their contents.

Integrity: Ensure the database's contents satisfy constraints.

Manipulation: Programming interface for accessing and modifying a database's contents.

RELATIONAL MODEL

A relation is an unordered set that contain the relationship of attributes that represent entities.

A tuple is a set of attribute values (also known as its domain) in the relation.
→ Values are (normally) atomic/scalar.
→ The special value **NULL** is a member of every domain (if allowed).

Artist(name, year, country)

name	year	country
Wu-Tang Clan	1992	USA
Notorious BIG	1992	USA
GZA	1990	USA

n-ary Relation
=
Table with *n* columns

RELATIONAL MODEL: PRIMARY KEYS

A relation's primary key uniquely identifies a single tuple.

Some DBMSs automatically create an internal primary key if a table does not define one.

DBMS can auto-generation unique primary keys via an identity column:

→ **IDENTITY** (SQL Standard?)
→ **SEQUENCE** (PostgreSQL / Oracle)
→ **AUTO_INCREMENT** (MySQL)

Artist(name, year, country)

name	year	country
Wu-Tang Clan	1992	USA
Notorious BIG	1992	USA
GZA	1990	USA

RELATIONAL MODEL: PRIMARY KEYS

A relation's primary key uniquely identifies a single tuple.

Some DBMSs automatically create an internal primary key if a table does not define one.

DBMS can auto-generation unique primary keys via an identity column:

→ **IDENTITY** (SQL Standard?)
→ **SEQUENCE** (PostgreSQL / Oracle)
→ **AUTO_INCREMENT** (MySQL)

Artist(id, name, year, country)

id	name	year	country
101	Wu-Tang Clan	1992	USA
102	Notorious BIG	1992	USA
103	GZA	1990	USA

RELATIONAL MODEL: FOREIGN KEYS

A foreign key specifies that an attribute from one relation maps to a tuple in another relation.

RELATIONAL MODEL: FOREIGN KEYS

Artist(id, name, year, country)

id	name	year	country
101	Wu-Tang Clan	1992	USA
102	Notorious BIG	1992	USA
103	GZA	1990	USA

Album(id, name, artists, year)

id	name	artists	year
11	Enter the Wu-Tang	101	1993
22	St.Ides Mix Tape	???	1994
33	Liquid Swords	103	1995

RELATIONAL MODEL: FOREIGN KEYS

Artist(id, name, year, country)

id	name	year	country
101	Wu-Tang Clan	1992	USA
102	Notorious BIG	1992	USA
103	GZA	1990	USA

ArtistAlbum(artist_id, album_id)

artist_id	album_id
101	11
101	22
103	22
102	22

Album(id, name, artists, year)

id	name	artists	year
11	Enter the Wu-Tang	101	1993
22	St.Ides Mix Tape	???	1994
33	Liquid Swords	103	1995

RELATIONAL MODEL: FOREIGN KEYS

ArtistAlbum(<u>artist_id</u>, <u>album_id</u>)

artist_id	album_id
101	11
101	22
103	22
102	22

Artist(<u>id</u>, name, year, country)

id	name	year	country
101	Wu-Tang Clan	1992	USA
102	Notorious BIG	1992	USA
103	GZA	1990	USA

Album(<u>id</u>, name, year)

id	name	year
11	Enter the Wu-Tang	1993
22	St.Ides Mix Tape	1994
33	Liquid Swords	1995

RELATIONAL MODEL: CONSTRAINTS

User-defined conditions that must hold for any instance of the database.

→ Can validate data within a single tuple or across entire relation(s).

→ DBMS prevents modifications that violate any constraint.

Unique key and referential (fkey) constraints are the most common.

SQL:92 supports global asserts but these are rarely used (too slow).

Artist(id, name, year, country)

id	name	year	country
101	Wu-Tang Clan	1992	USA
102	Notorious BIG	1992	USA
103	GZA	1990	USA

```
CREATE ASSERTION myAssert
 CHECK ( <SQL> );
```

DATA MANIPULATION LANGUAGES (DML)

Methods to store and retrieve information from a database.

Procedural:

→ The query specifies the (high-level) strategy to find the desired result based on sets / bags.

← Relational Algebra

Non-Procedural (Declarative):

→ The query specifies only what data is wanted and not how to find it.

← Relational Calculus

RELATIONAL ALGEBRA

Fundamental operations to retrieve and manipulate tuples in a relation.

→ Based on set algebra (unordered lists with no duplicates).

Each operator takes one or more relations as its inputs and outputs a new relation.

→ We can "chain" operators together to create more complex operations.

σ	Select
π	Projection
$\cup$	Union
$\cap$	Intersection
$-$	Difference
$\times$	Product
$\bowtie$	Join

RELATIONAL ALGEBRA: SELECT

Choose a subset of the tuples from a relation that satisfies a selection predicate.

→ Predicate acts as a filter to retain only tuples that fulfill its qualifying requirement.

→ Can combine multiple predicates using conjunctions / disjunctions.

Syntax: $\sigma_{predicate}(R)$

R(a_id,b_id)

a_id	b_id
a1	101
a2	102
a2	103
a3	104

$\sigma_{a\_id='a2'}(R)$

a_id	b_id
a2	102
a2	103

$\sigma_{a\_id='a2' \land b\_id>102}(R)$

a_id	b_id
a2	103

```
SELECT * FROM R
 WHERE a_id='a2' AND b_id>102;
```

RELATIONAL ALGEBRA: PROJECTION

Generate a relation with tuples that contains only the specified attributes.
- → Rearrange attributes' ordering.
- → Remove unwanted attributes.
- → Manipulate values to create derived attributes.

Syntax: $\Pi_{A1,A2,\ldots,An}(R)$

R(a_id,b_id)

a_id	b_id
a1	101
a2	102
a2	103
a3	104

$\Pi_{b\_id-100,a\_id}(\sigma_{a\_id='a2'}(R))$

b_id-100	a_id
2	a2
3	a2

```
SELECT b_id-100, a_id
  FROM R WHERE a_id = 'a2';
```

RELATIONAL ALGEBRA: UNION

Generate a relation that contains all tuples that appear in either only one or both input relations.

Syntax: (R ∪ S)

R(a_id,b_id)

a_id	b_id
a1	101
a2	102
a3	103

S(a_id,b_id)

a_id	b_id
a3	103
a4	104
a5	105

```
(SELECT * FROM R)
UNION
(SELECT * FROM S);
```

(R ∪ S)

a_id	b_id
a1	101
a2	102
a3	103
a4	104
a5	105

RELATIONAL ALGEBRA: INTERSECTION

Generate a relation that contains only the tuples that appear in both of the input relations.

Syntax: (R ∩ S)

R(a_id,b_id)

a_id	b_id
a1	101
a2	102
a3	103

S(a_id,b_id)

a_id	b_id
a3	103
a4	104
a5	105

(R ∩ S)

a_id	b_id
a3	103

```
(SELECT * FROM R)
INTERSECT
(SELECT * FROM S);
```

RELATIONAL ALGEBRA: DIFFERENCE

Generate a relation that contains only the tuples that appear in the first and not the second of the input relations.

Syntax: (R - S)

R(a_id,b_id)

a_id	b_id
a1	101
a2	102
a3	103

S(a_id,b_id)

a_id	b_id
a3	103
a4	104
a5	105

(R - S)

a_id	b_id
a1	101
a2	102

```
(SELECT * FROM R)
EXCEPT
(SELECT * FROM S);
```

RELATIONAL ALGEBRA: PRODUCT

Generate a relation that contains all possible combinations of tuples from the input relations.

Syntax: (R × S)

```
SELECT * FROM R CROSS JOIN S;
```

```
SELECT * FROM R, S;
```

R(a_id,b_id)

a_id	b_id
a1	101
a2	102
a3	103

S(a_id,b_id)

a_id	b_id
a3	103
a4	104
a5	105

(R × S)

R.a_id	R.b_id	S.a_id	S.b_id
a1	101	a3	103
a1	101	a4	104
a1	101	a5	105
a2	102	a3	103
a2	102	a4	104
a2	102	a5	105
a3	103	a3	103
a3	103	a4	104
a3	103	a5	105

RELATIONAL ALGEBRA: JOIN

Generate a relation that contains all tuples that are a combination of two tuples (one from each input relation) with a common value(s) for one or more attributes.

Syntax: $(R \bowtie S)$

R(a_id,b_id)

a_id	b_id
a1	101
a2	102
a3	103

S(a_id,b_id,val)

a_id	b_id	val
a3	103	XXX
a4	104	YYY
a5	105	ZZZ

R.a_id	R.b_id	~~S.a_id~~	~~S.b_id~~	S.val
a3	103	~~a3~~	~~103~~	XXX

$(R \bowtie S)$

a_id	b_id	val
a3	103	XXX

RELATIONAL ALGEBRA: JOIN

Generate a relation that contains all tuples that are a combination of two tuples (one from each input relation) with a common value(s) for one or more attributes.

Syntax: $(R \bowtie S)$

R(a_id,b_id)

a_id	b_id
a1	101
a2	102
a3	103

S(a_id,b_id,val)

a_id	b_id	val
a3	103	XXX
a4	104	YYY
a5	105	ZZZ

$(R \bowtie S)$

a_id	b_id	val
a3	103	XXX

```
SELECT * FROM R NATURAL JOIN S;
```

```
SELECT * FROM R JOIN S USING (a_id, b_id);
```

```
SELECT * FROM R JOIN S
    ON R.a_id = S.a_id AND R.b_id = S.b_id;
```

RELATIONAL ALGEBRA: EXTRA OPERATORS

Rename (ρ)

Assignment ($R \leftarrow S$)

Duplicate Elimination (δ)

Aggregation (γ)

Sorting (τ)

Division ($R \div S$)

OBSERVATION

Relational algebra defines an ordering of the high-level steps of how to compute a query.

→ Example: $\sigma_{b\_id=102}(R \bowtie S)$ vs. $(R \bowtie (\sigma_{b\_id=102}(S))$

A better approach is to state the high-level answer that you want the DBMS to compute.

→ Example: Retrieve the joined tuples from **R** and **S** where **b_id** equals 102.

RELATIONAL MODEL: QUERIES

The relational model is independent of any query language implementation.

SQL is the *de facto* standard (many dialects).

```
for line in file.readlines():
  record = parse(line)
  if record[0] == "GZA":
    print(int(record[1]))
```

```
SELECT year FROM artists
 WHERE name = 'GZA';
```

DATA MODELS

Relational

Key/Value

Graph

Document / XML / Object ← Leading Alternative

Wide-Column / Column-family

Array / Matrix / Vectors ← Current Hotness

Hierarchical

Network

Multi-Value

DOCUMENT DATA MODEL

A collection of record documents containing a hierarchy of named field/value pairs.

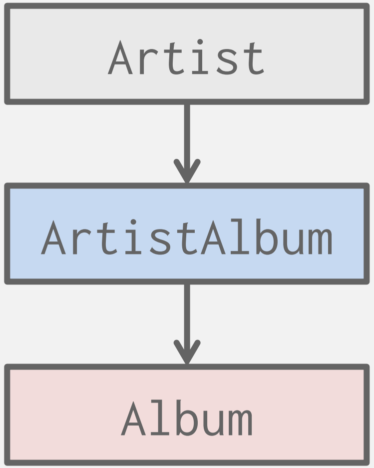
→ A field's value can either a scalar type, an array of values, or another document.

→ Modern implementations use JSON. Older systems use XML or custom object representations.

Avoid "relational-object impedance mismatch" by tightly coupling objects and database.

MongoDB RAVENDB amazon DynamoDB Firebase Couchbase MarkLogic fauna

DOCUMENT DATA MODEL



→ R_1(id,…)

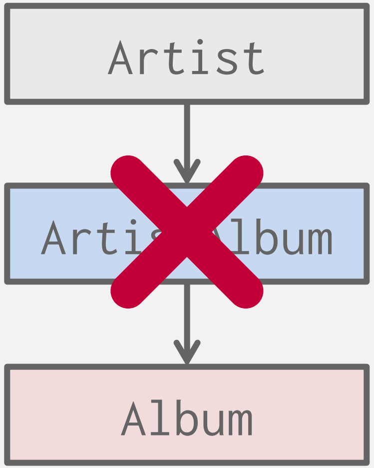
$\bowtie$

→ R_2(artist_id,album_id)

$\bowtie$

→ R_3(id,…)

DOCUMENT DATA MODEL



→ R_1(id,…)

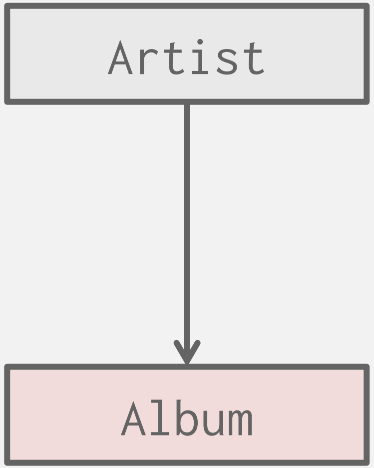
⋈

→ R_2(artist_id,album_id)

⋈

→ R_3(id,…)

DOCUMENT DATA MODEL



Application Code

```
class Artist {
    int id;
    String name;
    int year;
    Album albums[];
}
class Album {
   int id;
   String name;
   int year;
}
```

```
{
 "name": "GZA",
 "year": 1990,
 "albums": [
   {
    "name": "Liquid Swords",
    "year": 1995
   },
   {
    "name": "Beneath the Surface",
    "year": 1999
  }
 ]
}
```

VECTOR DATA MODEL

One-dimensional arrays used for nearest-neighbor search (exact or approximate).

→ Used for semantic search on embeddings generated by ML-trained transformer models (think ChatGPT).

→ Native integration with modern ML tools and APIs (e.g., LangChain, OpenAI).

At their core, these systems use specialized indexes to perform NN searches quickly.

Pinecone Weaviate milvus qdrant marqo LanceDB featureform

VECTOR DATA MODEL

One-dimensional arrays used for nearest-neighbor search (exact or approximate).

→ Used for semantic search on embeddings generated by ML-trained transformer models (think ChatGPT).

→ Native integration with modern ML tools and APIs (e.g., LangChain, OpenAI).

At their core, these systems use specialized indexes to perform NN searches quickly.

Pinecone Weaviate milvus qdrant

marqo LanceDB featureform

VECTOR DATA MODEL

Album(id, name, year)

id	name	year
11	Enter the Wu-Tang	1993
22	St.Ides Mix Tape	1994
33	Liquid Swords	1995

OpenAI Hugging Face

Transformer

Embeddings

Id1 → [0.32, 0.78, 0.30, ...]
Id2 → [0.99, 0.19, 0.81, ...]
Id3 → [0.01, 0.18, 0.85, ...]
⋮

Vector Index

HNSW, IVFFlat
Meta Faiss, Spotify Annoy

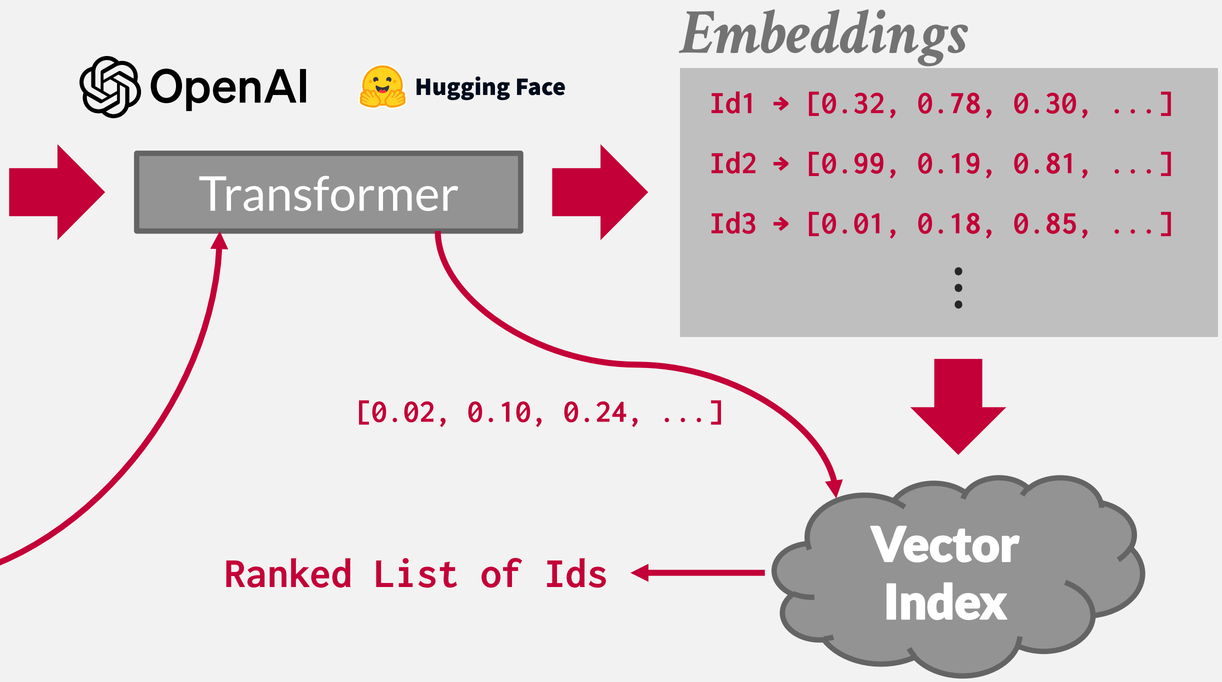
VECTOR DATA MODEL

Album(id, name, year)

id	name	year
11	Enter the Wu-Tang	1993
22	St.Ides Mix Tape	1994
33	Liquid Swords	1995

Query

Find albums similar to "Liquid Swords"



HNSW, IVFFlat
Meta Faiss, Spotify Annoy

CONCLUSION

Databases are ubiquitous.

Relational algebra defines the primitives for processing queries on a relational database.

We will see relational algebra again when we talk about query optimization + execution.

NEXT CLASS

Modern SQL
→ Make sure you understand basic SQL before the lecture.