Relational Data Model

Types of Data

- Structured data
  - CSV tables
  - The largest category on kaggle.com
- Semi-structured data
  - JSON files
- Unstructured data
  - Text, web pages

Data Model

- Structure
- Values constraints
- Operations

Relational Data Model

- Data model for CSV tables
- Structure
  - TABLE or RELATION is the only element
- Value constraints
  - Unique or keys
  - NULLs
- Operations
  - Relational algebra or algebra for tables

TABLE Or Relation (1)

- Attributes or columns
  - Table header: name, latitude, longitude
  - Type or domain
    - Primitive: int, float, char[], string or varchar[]
    - Containers not allowed
- Schema
  - Cal_Cities (name, latitude, longitude)
- Tuples
  - (Merced, 37.302164, -120.482967)

TABLE Or Relation (2)

- Simple and general
  - (Any) Type of data can be represented as a table
- Abstract representation from implementation
  - Array (vector) of struct
  - Linked list of struct
  - Hash table of struct

Keys and NULLs

- Key
  - Attribute (or set of attributes) that have unique (different) values across all the tuples
  - There are no two different tuples which have the same value for the key attribute
  - Cal_Cities → name
- NULL
  - Missing value for an attribute in a tuple
  - Cal_Cities_Pop → pop_1980

Relational Algebra

- Set of operations on tables
  - A table is seen as a collection (or set) of tuples
  - Cannot index in the table
    - Cal_Cities[7] is not a valid operation
- Single table operations
  - Select column, select tuple (row), aggregate, grouping
- Multiple table operations
  - Product and Join, Union, Intersection, Difference

Schema Examples

- California_Cities
- Computers
- TPCH