Data Engineering and Data Models

Data Science vs. Data Engineering

Data Science

- Building models
- Answering business questions
- Uses Statistics / Machine learning
- Expects data to be available and computation to be fast.

Data Engineering

- Builds the data and computation infrastructure used by the data scientist.
- Uses relational databases, streaming, cloud computing …

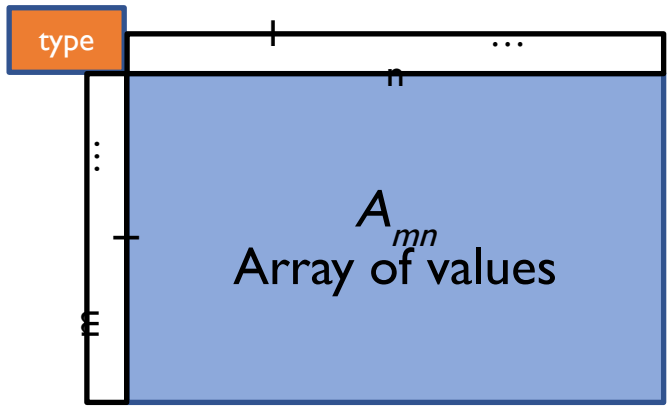
What data engineering we will cover

- **Will Cover**
  - Hadoop File System
  - Data partitioning and partition balancing
  - Caching and persistence
  - Checkpointing
- **Will Not Cover**
  - Data Cleaning
  - Spark Server configuration and optimization.
  - Creating scalable pipelines
  - Containerization

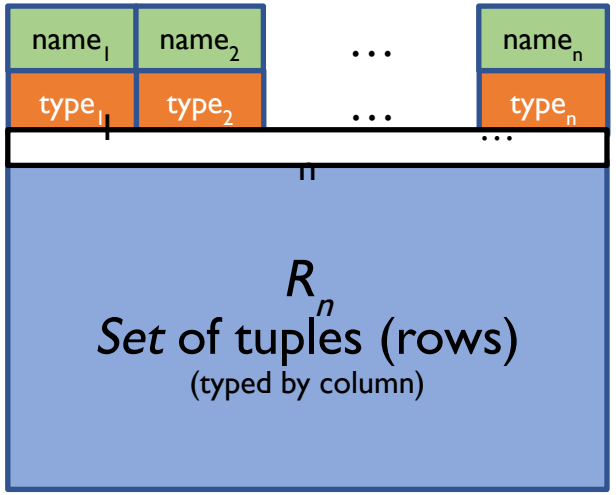
Three Data Models: Relations, Tensors and Dataframes

Data Scientists and Data Engineering communicate using these data models

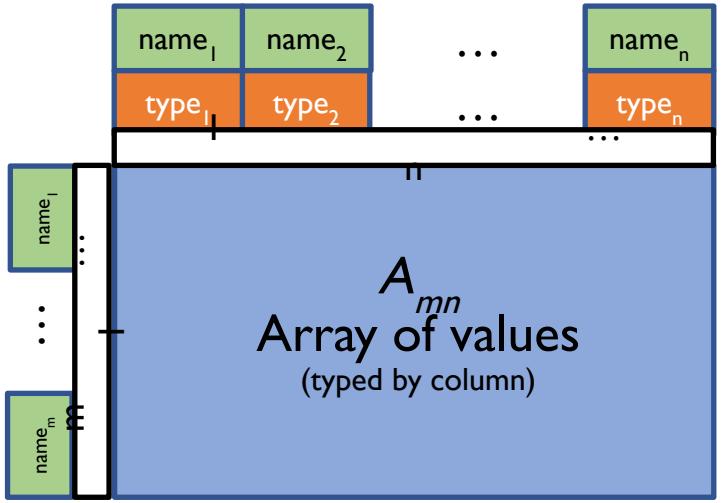
Three Data Models (and languages)



Matrix



Relation/
Table



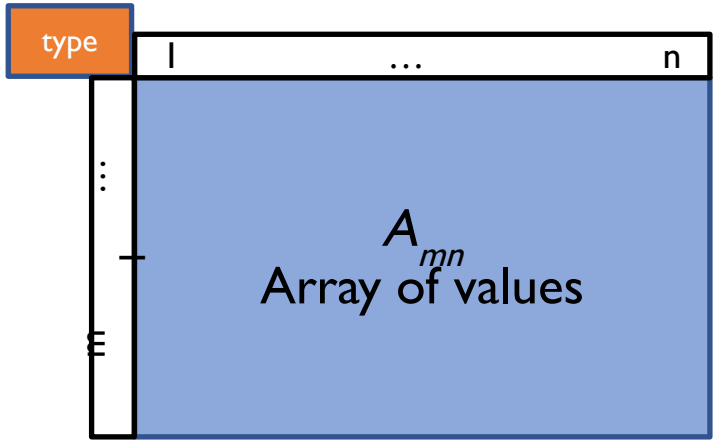
Dataframe

Linear
Algebra

Relational
Algebra

?

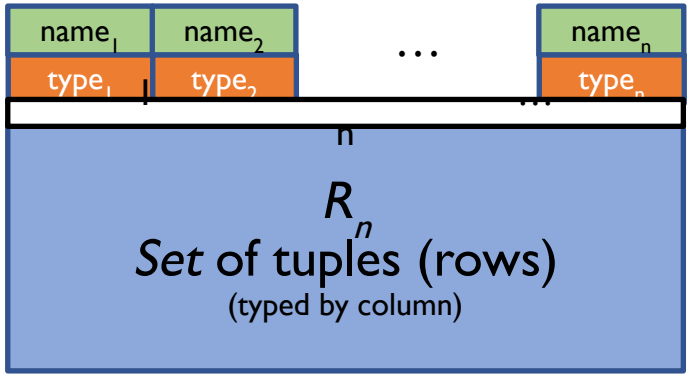
Matrices



Matrix

- A Rectangle of numbers
- All numbers are of the same type
- Can transpose (exchange rows and columnns)
- Can add and multiply.
- Typically small enough to reside in memory of one computer.

Tables



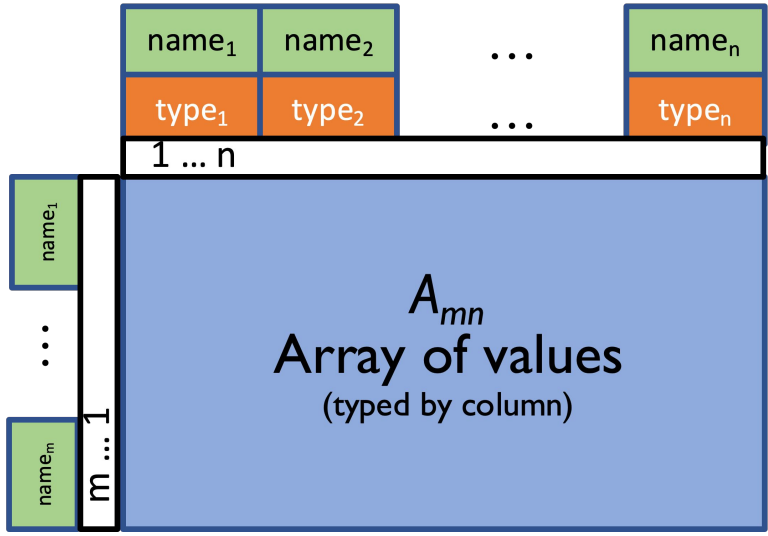
Table

- **Row:** A tuple – defines an entity
- **Column:** a named property of the entity: each column has a type
- **Schema**: a collection of related tables.
- **Keys:** a special column (1 per table) that uniquely identifies
- Can select a set of rows based on a condition **(SQL)**
- Can be very large (TB) and reside on disk.
- Disk Data Structures make retrieval much faster than flat files.

Why Use DataBases

- We can do the same in python/C/Matlab
- But only if they fit in memory
- A Database can span many disks on many computers.
- Disk Data Structures to make retrieval much faster than flat files.

Dataframes



- A blend of ideas from relations and matrices
  - Originally defined in the S language, which led to R
  - Ordered, named rows and columns.
  - But only columns have types.
- Beware: there is not a standard definition of dataframes
  - Some define dataframes as relations with an ordering key. *No transpose!*
- Two popular flavors: Pandas DataFrames and Spark DataFrames.

Summary

- Data Science: Data analytics on large complex data.
- Data Engineering: making the analytics scale on very large data.
- Matrices are the basic data structure for linear algebra, Neural Networks.
- Tables are the basic data structure for relational databases.
- DataFrames are a compromise between tables and matrices.