Cross-language Data Development with Apache Arrow

Brief Introduction

👋🤓 Hi, I’m Dave

University of Colorado Anschutz Medical Campus
Department of Biomedical Informatics
Software Engineering Team

Presentation Outline

1. ✍️ Data Literacy, Data Grammar, and Software Diversity
2. 📚 Apache Arrow Concepts
3. ⏯️ Examples

Preface

Why does this matter?

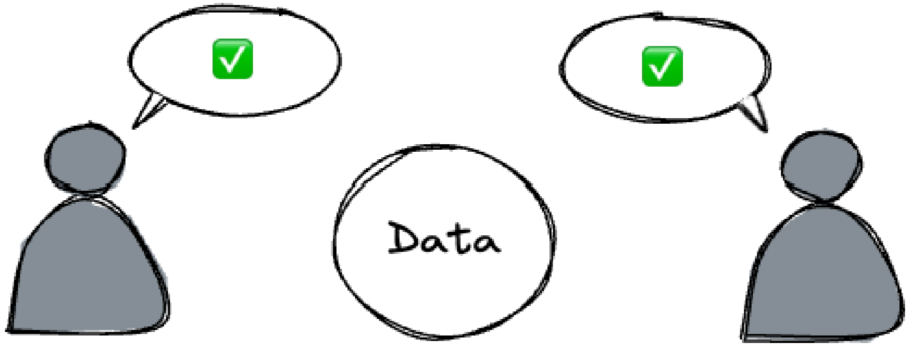
- Data is locked up by technology and language differences.
- Sometimes this happens by accident or for performance.
- How can you free your data to create opportunity?

Data Literacy

Data Literacy (Wikipedia)

> *“Data literacy is the ability to read, understand, create, and communicate data as information.”*

Data Literacy



How we might imagine data conversations.

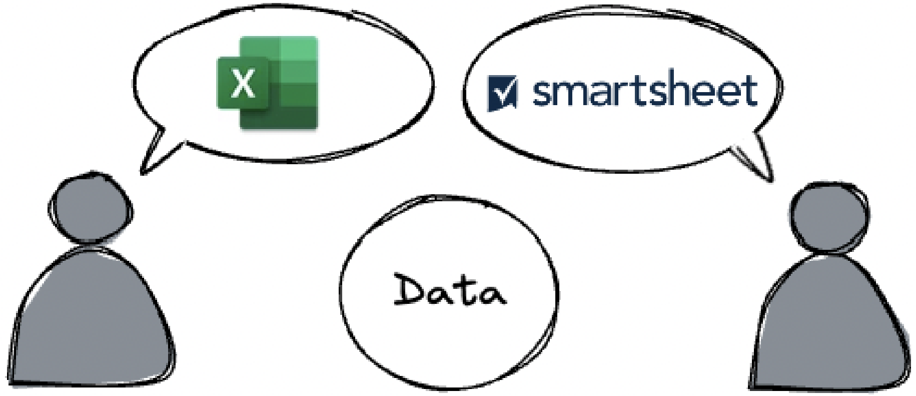
Data Literacy

A CSV file:

```
a_string , b_int , c_float
"dog"    , 1     , NaN
'cat'    , NULL  ,  "0.2f"
         , 3     ,  0.8
```

What data type are these columns (strings, floats, integers)?

Data Literacy



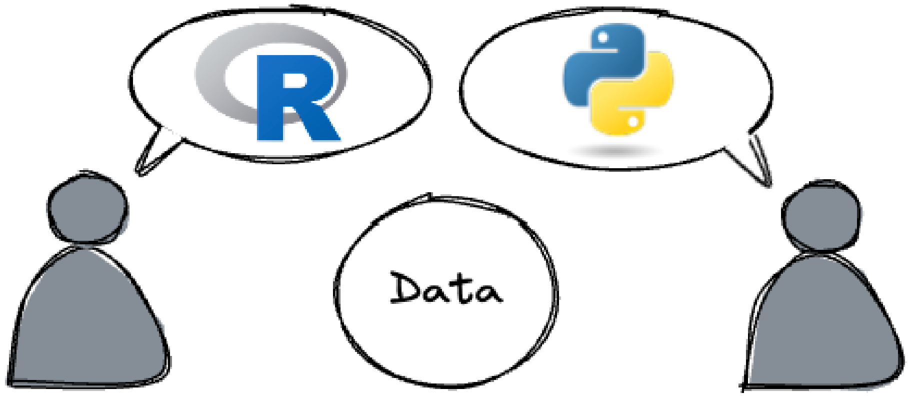
The spreadsheets should be the same, right?

Data Literacy

What is data, ***really***?

- What is a “datatype”?
- What is a “table”?
- What is a “schema”?
- What is a “dataframe”?

Data Literacy



We can develop our way around this!

Data Literacy

R data.table

```r
library(data.table)

DT = as.data.table(iris)

DT[Petal.Width > 1.0,
   mean(Petal.Length),
   by = Species]
```

Python Pandas.DataFrame

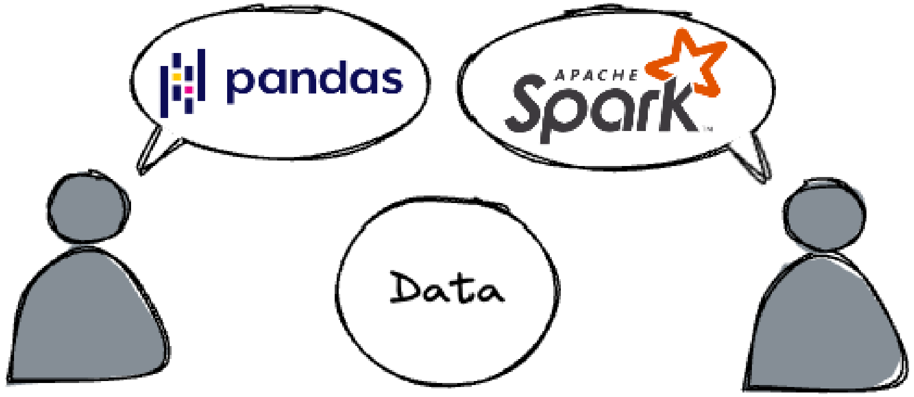
```python
import pandas as pd

df = pd.read_csv(...)

df[df["Petal.Width"] > 1.0].groupby(
    "Species"
)["Petal.Length"].mean()
```

How different could R and Python be?

Data Literacy

Maybe it gets better if we choose one language?

Data Literacy

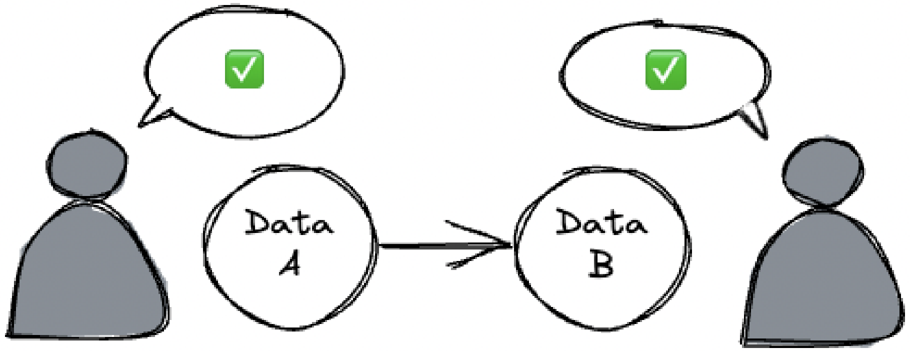


They’re all just dataframes, right?

Data Literacy

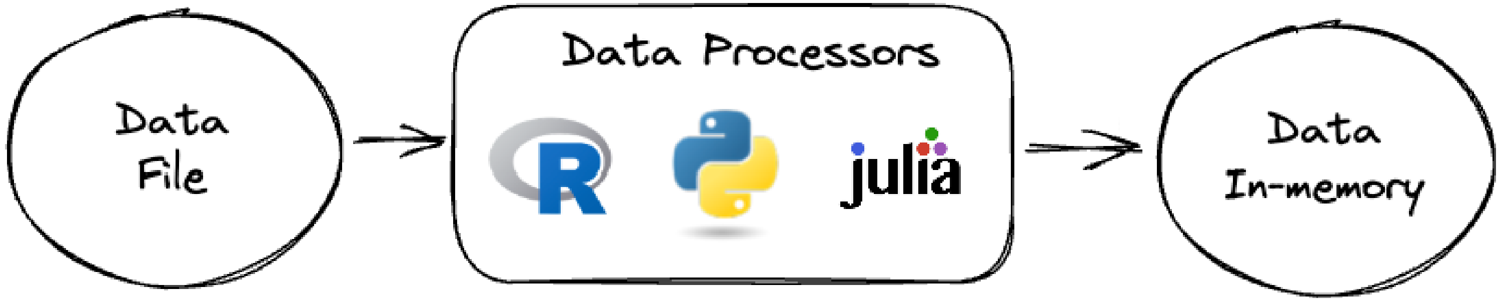
Which data approach is more ***"correct"***?

Data Grammar



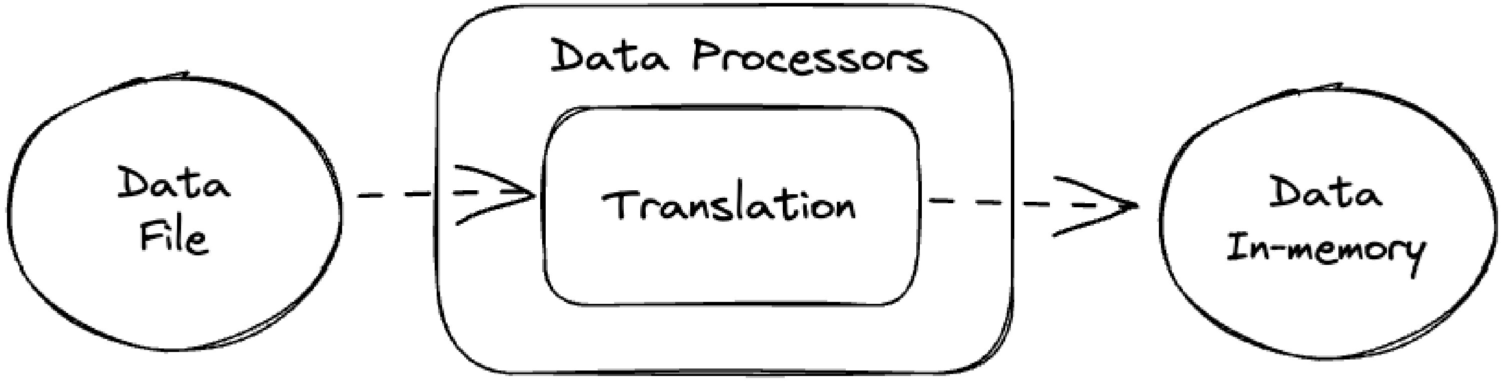
In addition to understanding what data is (**literacy**), we need ways to use the data too (**grammar**).

Data Grammar



Data in-memory isn't the same as data in a file.

Data Grammar



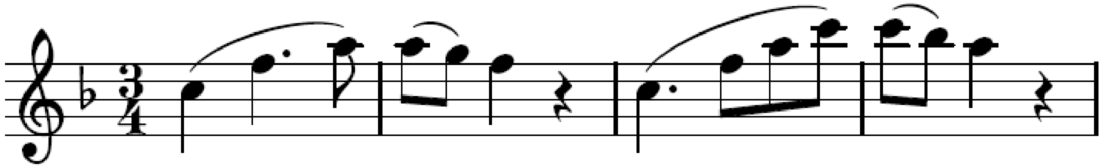
Each processor uses opinionated translations.

Data Grammar

- Each translation without a common grammar is different.
- How much can we hope to understand one another?

Data Grammar

A quick analogy:



Music notes and how they are played together.

Image from Wikimedia Commons: Public Domain

Data Grammar

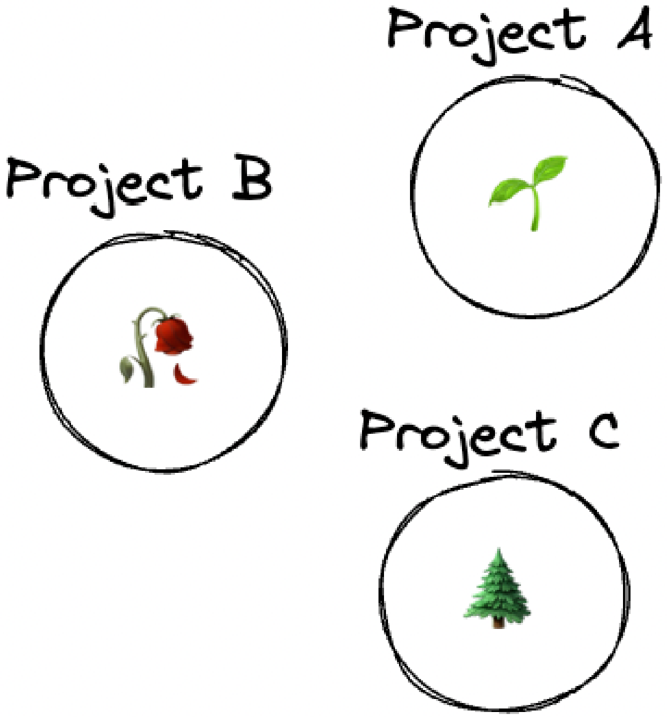
- What kind of data “music” do you play?
- How does your “band” play together?

Software Diversity

🪴 **Software gardening:**

A practice of growing and cultivating software using parallels from horticulture.

Software Diversity

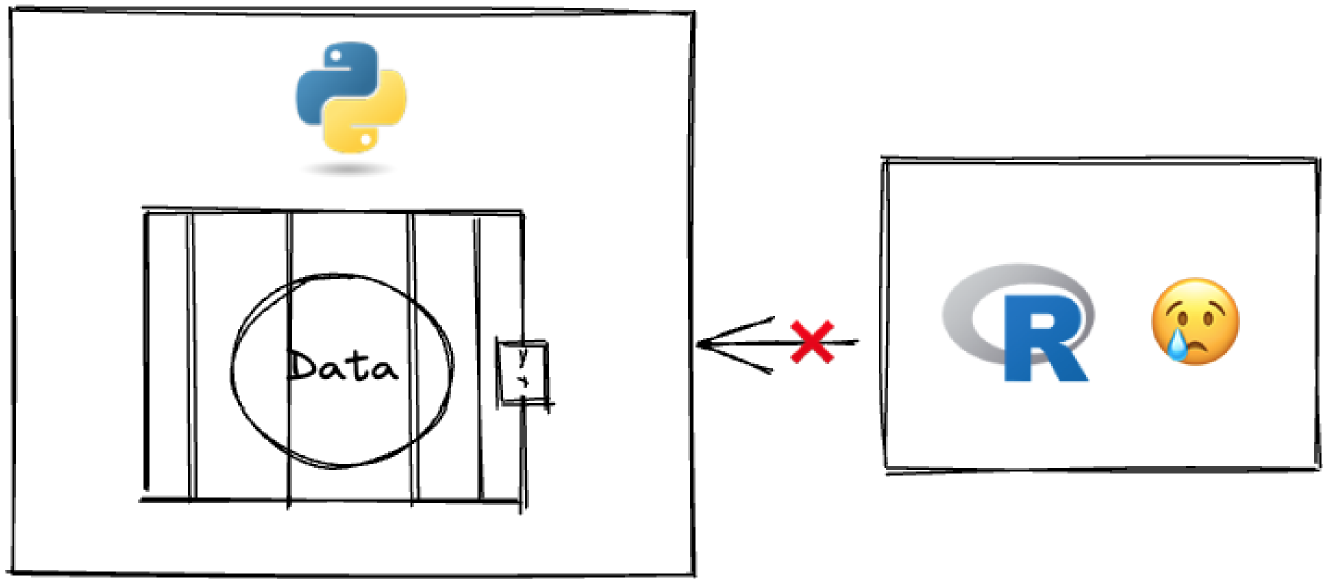


Software can follow patterns from life.

Software Diversity

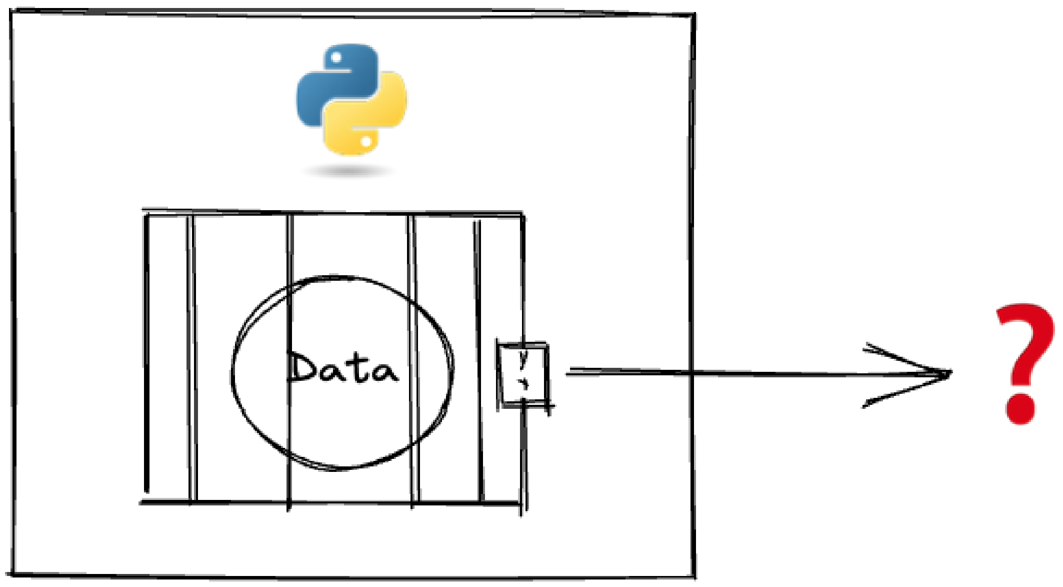
Time influences software.

Software Diversity



Single-stack or mono-lingual restrictions for your ecosystem mean isolation.

Software Diversity



Isolation may mean lower chances of survival (what’s next?).

Software Diversity

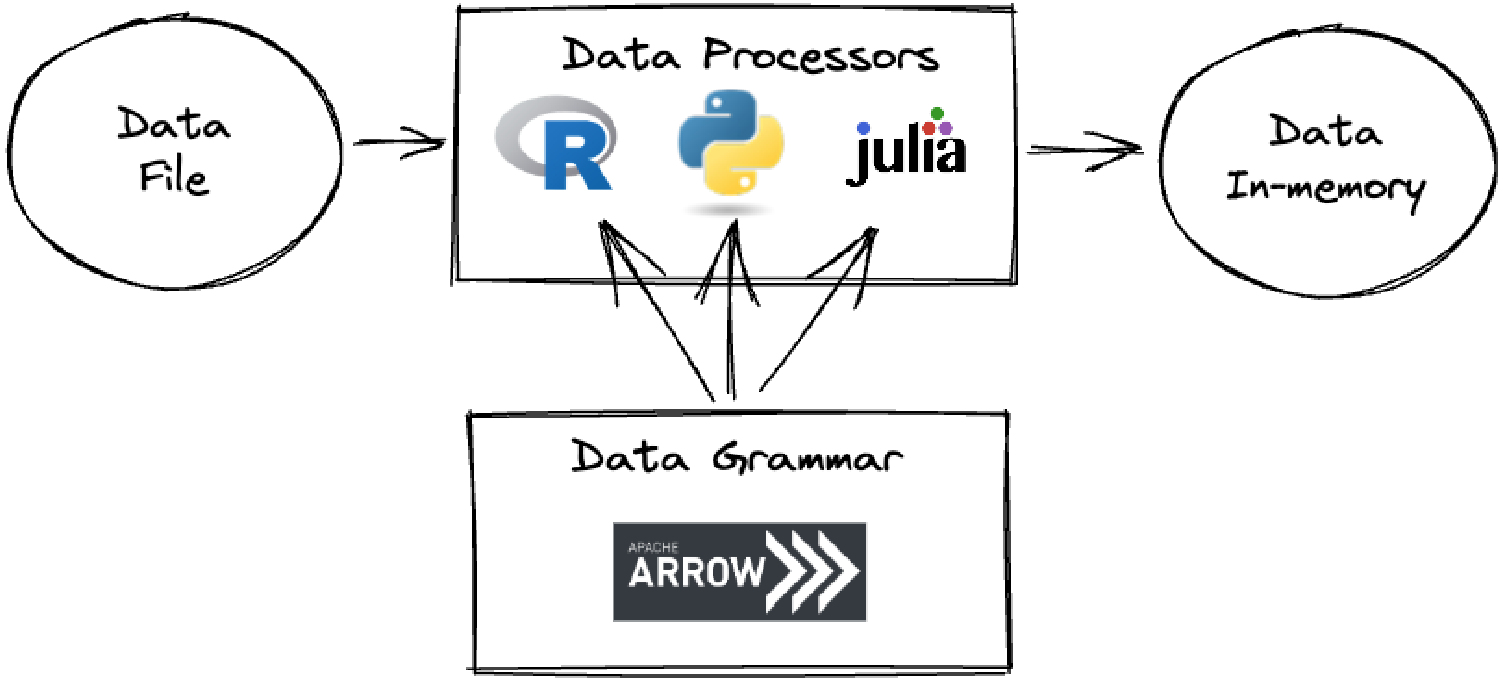
- How can software diversity and common data grammar be handled together?
- Isn’t this all contradictory?

Apache Arrow

Apache Arrow is a library for processing data across many languages.

(https://arrow.apache.org)

Apache Arrow



Arrow enables a data grammar for software diversity.

Apache Arrow

Arrow’s key features:

- **Language interoperability**
  (bindings for R, Python, Julia, Java, more…)
- **Metadata compatibility and availability**
  (types, schema, descriptions, more…)
- **Performance and “zero-copy” capabilities**
  (shared memory buffers, avoid conversions)

Apache Arrow

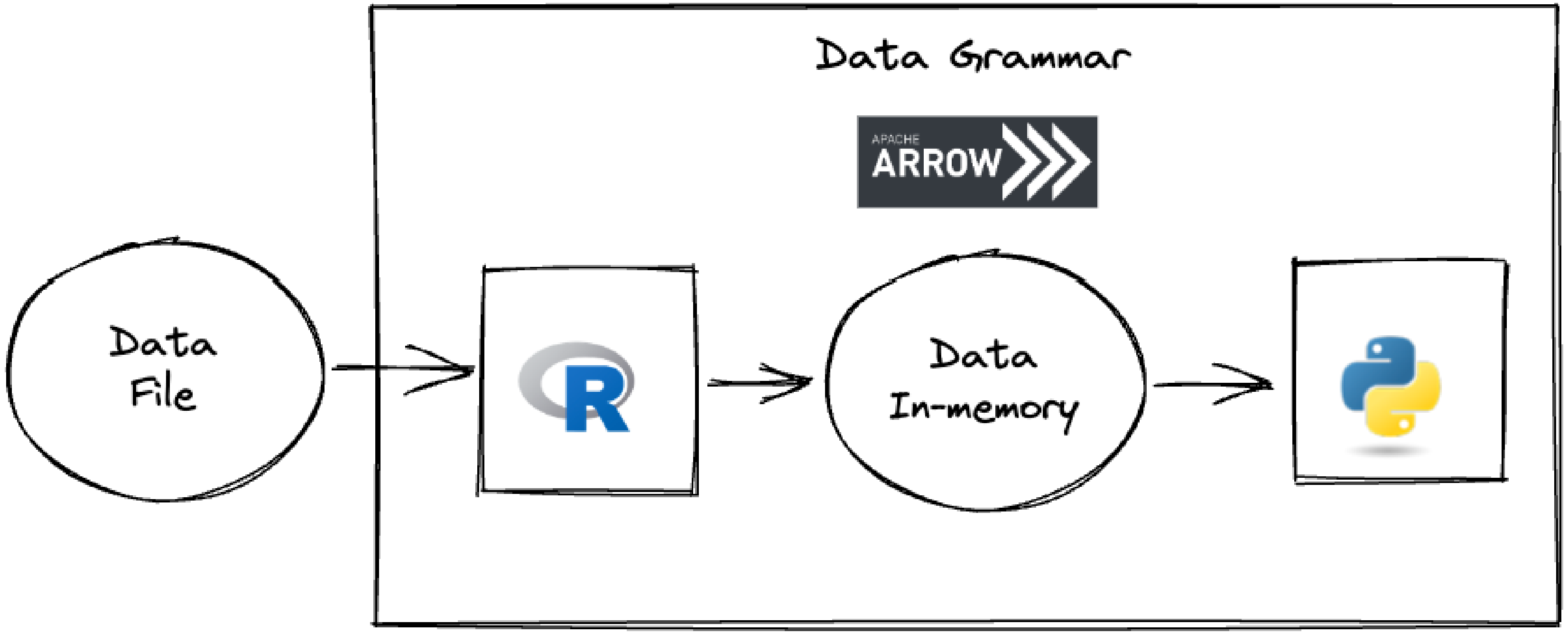
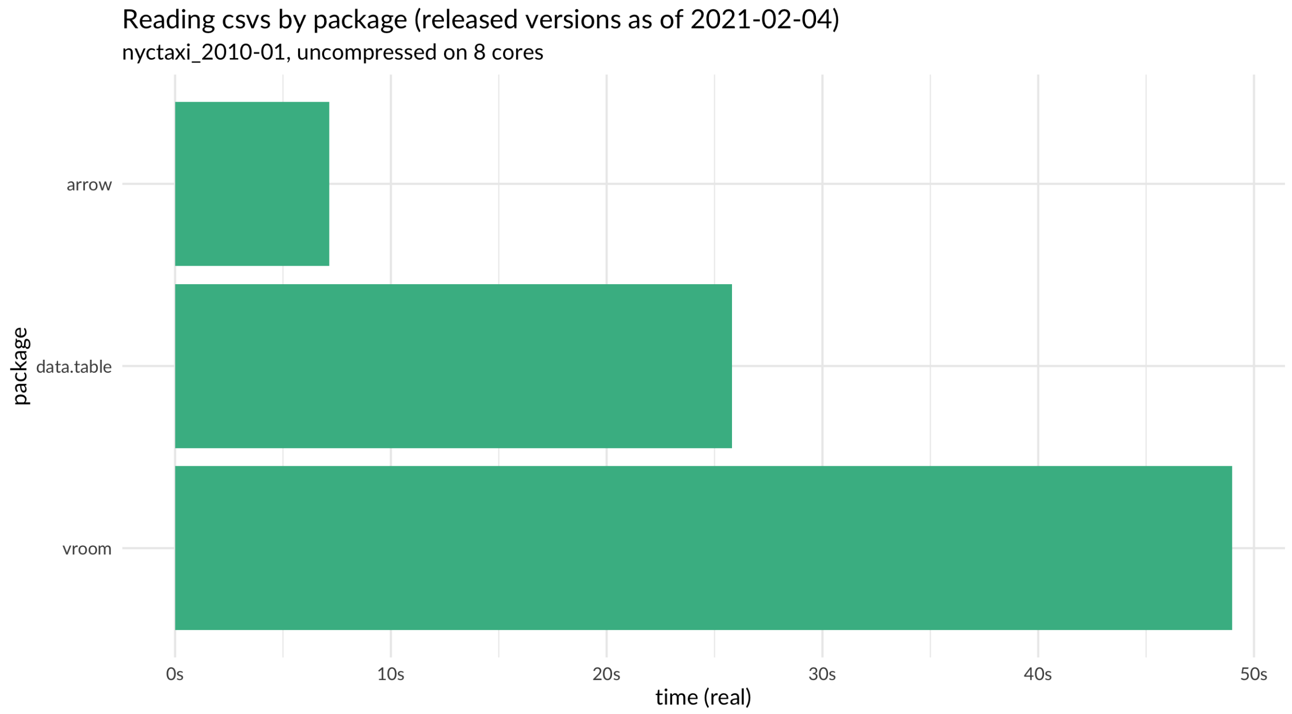


Illustration of “zero-copy” at work between R and Python.

Apache Arrow



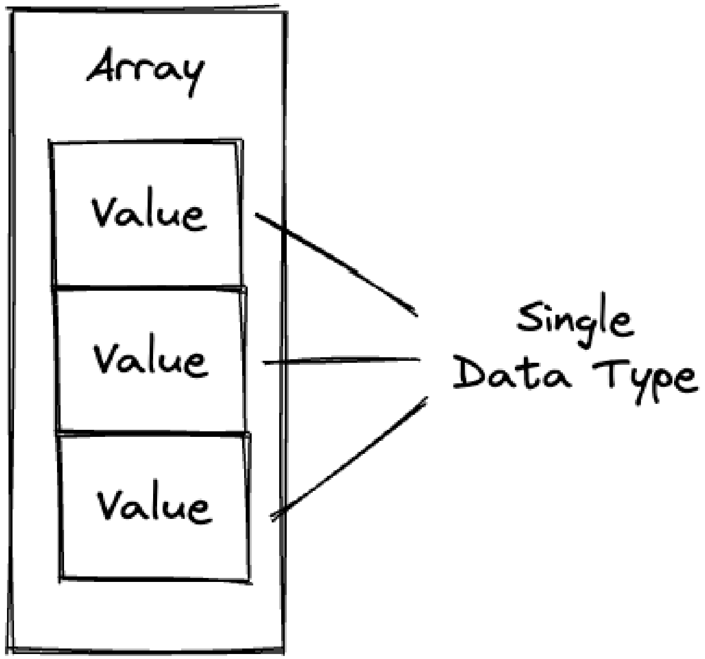
Performance is another reason to make use of Arrow.

Chart from: Ursa Labs (Voltron Data), Measuring and Monitoring Arrow's Performance: Some Updated R Benchmarks

Apache Arrow - Concepts

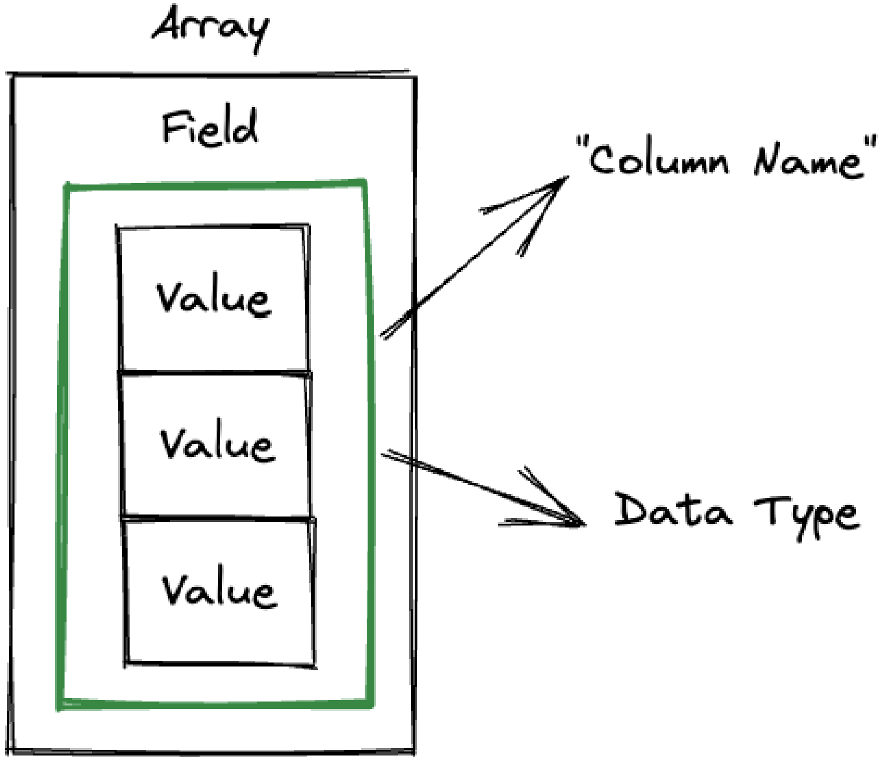
Covering a few brief concepts (there’s much more!).

Apache Arrow - Concepts



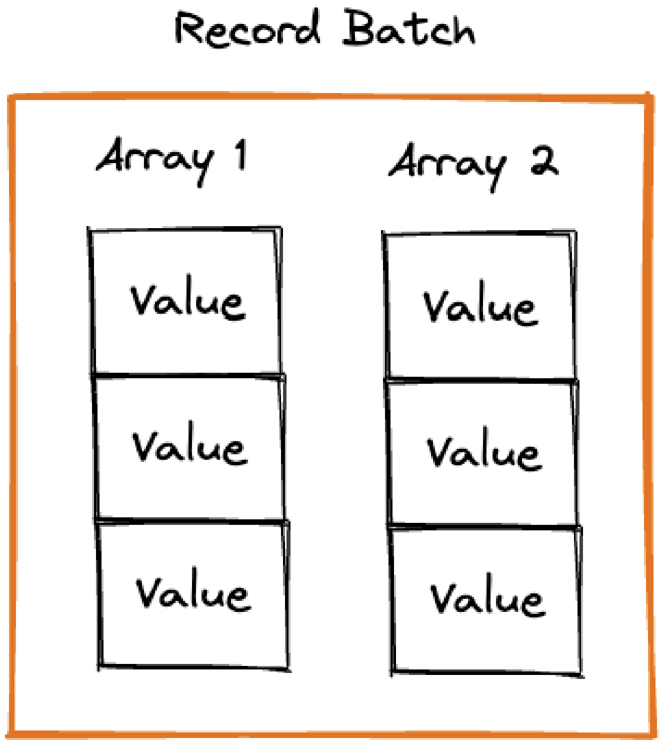
Arrays are “columns”: they include values of a single type.

Apache Arrow



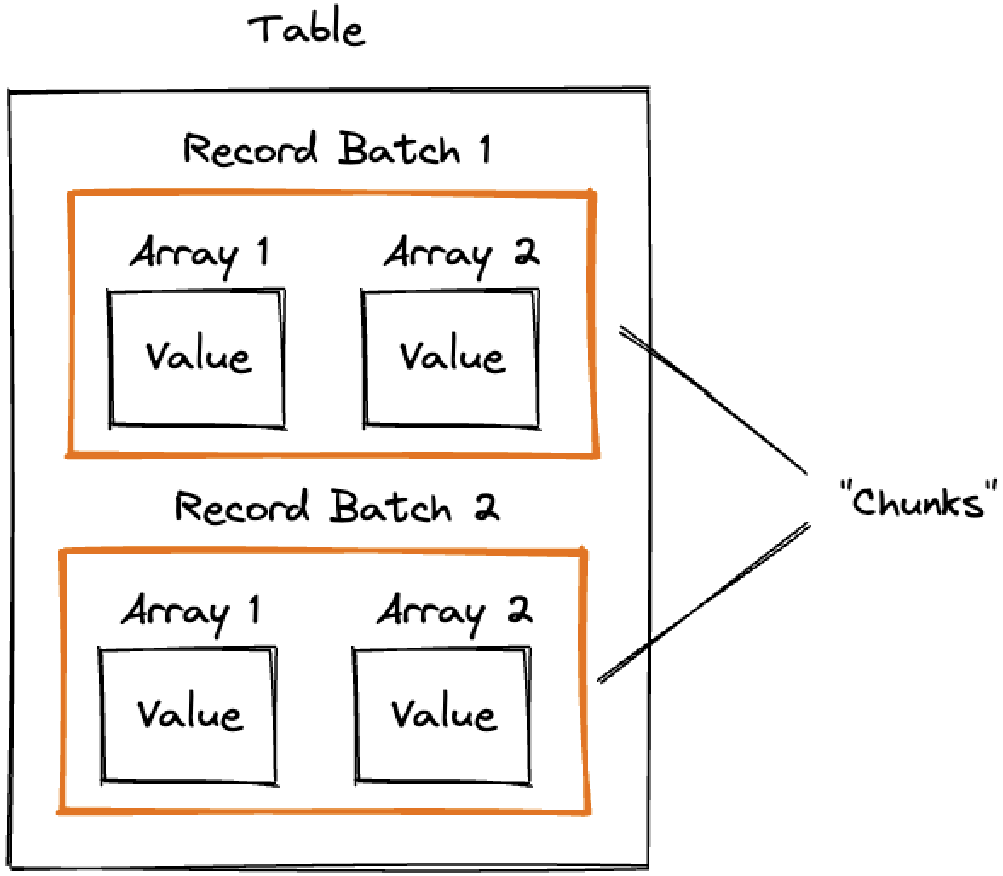
Array **Fields** may include a name, data type, and other metadata.

Apache Arrow



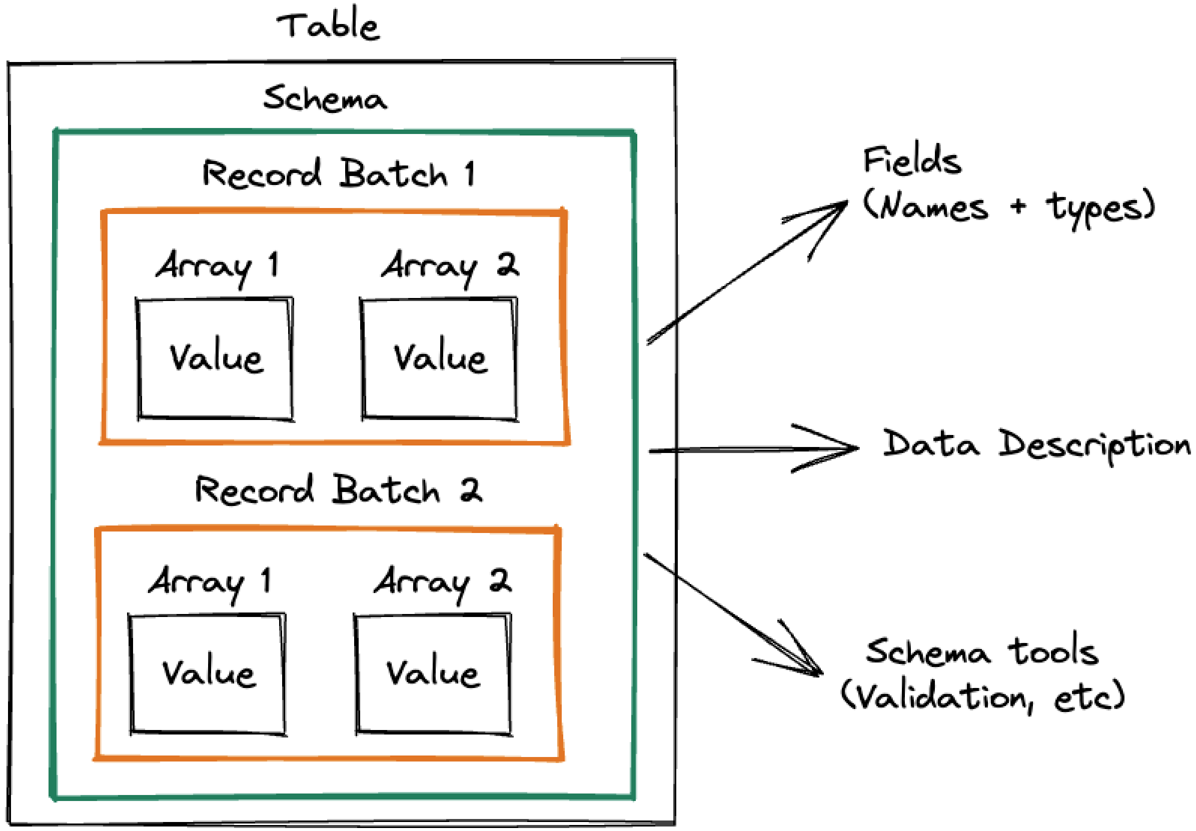
Record Batches are collections of arrays.

Apache Arrow



Tables are collections of Record Batches.

Apache Arrow



Tables include **Schema** which collect fields, data description, and metadata tools.

Apache Arrow - Examples

```r
library(dplyr)
library(arrow)

# read iris data into arrow table
iris_table <- arrow::arrow_table(iris)

# Use arrow and dplyr to form result
result <- iris_table %>%
  filter(Petal.Width > 1.0) %>%
  group_by(Species) %>%
  dplyr::summarize(mean_Petal_Length = mean(Petal.Length)) %>%
  # lazy evaluation
  collect()

# Print the result
print(result)
```

R with Arrow and Dplyr example.

Apache Arrow - Examples

```r
# Load the necessary packages
library(dplyr)
library(arrow)
library(reticulate)

# create pyarrow python environment
virtualenv_create("my-pyarrow-env")
use_virtualenv("my-pyarrow-env")
install_pyarrow("my-pyarrow-env")

# read iris data into arrow table
iris_table <- arrow::arrow_table(iris)

# print out the R-based arrow iris table
print(iris_table)

# send the R-based arrow iris table to pyarrow
pyarrow_table <- r_to_py(iris_table)
```

Opening up the PyArrow API via R.

Apache Arrow - Examples

```python
import duckdb
from pyarrow import csv

# read iris CSV data into arrow table
arrow_table = csv.read_csv("iris.csv")

# perform a SQL query on arrow table using duckdb
duckdb.connect().execute(
    """
    SELECT
        Species,
        AVG(Petal_Length) as mean_Petal_Length
    FROM arrow_table
    WHERE Petal_Width > 1.0
    GROUP BY Species
    """
).arrow()
```

Performing a SQL query on Arrow data in Python.

Thank you!

Questions / Comments?

Further References

- Arrow for R Cheatsheet
- PyArrow Documentation