Research in Progress

# Bridges for Tabular Dimension Chasms

# Presentation Outline

1. Tabular data formats
2. Dimensional data
3. Relationships

# Context

Why?

- CytoTable: a work in progress building on these concepts.

# Tabular data

| Col_A | Col_B | Col_C |
|---|---|---|
| 1 | a | 0.01 |
| 2 | b | 0.02 |

A visual example of tabular data.

# Tabular data

- **Definitions**:
  - **Encoding**: "convert into a "coded" form."
  - **Coded**: "converted into a code to convey a secret meaning."

Source: Oxford Languages via Google

# Tabular data

🤖 💬 🤖

Computer secrets? 🤷

# Tabular data • CSV

```
Col_A,Col_B,Col_C
1,a,0.01
2,b,0.02
```

A CSV (comma delimited spreadsheet) table.

# Tabular data · CSV

Strengths of CSV's

- Simple
- Interoperable
- Human-readable

# Tabular data • CSV

```
Col_A,Col B,Col_C,COL_D
,a,"0.01"
2,null,0.02,{'color':'blue'}
```

A challenging CSV table.

# Tabular data · CSV

Challenges with CSV's

- No data types
- Expensive to slice (cols or rows)
- Missing data handling
- 2D dimensionality

# Tabular data • Parquet

```
file.parquet (unable to view)
```

Parquet files stored as a table.

# Tabular data • Parquet

What even is a parquet file?

Why would we use it?

# Tabular data • Parquet

- "**Apache Parquet** is an open source, column-oriented data file format designed for efficient data storage and retrieval." (https://parquet.apache.org/)

# Tabular data · Parquet

- Column orientation?
  - Data is stored with column-wise access in mind.
  - Whereas, with CSV, we must access data row-wise.
  - We can access a single column without reading the full dataset.

# Tabular data • Parquet

CSV Reads

| Col_A | Col_B | Col_C |
|---|---|---|
| 1 | a | 0.01 |
| 2 | b | 0.02 |

# Tabular data • Parquet

Parquet Reads

| Col_A | Col_B | Col_C |
|---|---|---|
| 1 | a | 0.01 |
| 2 | b | 0.02 |

# Tabular data • Parquet

“file1.parquet”

| Col_A | Col_B | Col_C |
|---|---|---|
| 1 | a | 0.01 |

"file2.parquet

| Col_A | Col_B | Col_C |
|---|---|---|
| 2 | b | 0.02 |

“Parquet dataset”

| Col_A | Col_B | Col_C |
|---|---|---|
| 1 | a | 0.01 |
| 2 | b | 0.02 |

Parquet files can be "chunks" of rows from a directory.

(They all need the same columns + types to do this)

# Tabular data · Parquet

- Open question: what if columns or groups of columns were split into subdirectories?
- Some inspiration for this: Firebolt whitepaper

# Tabular data • Parquet

```
ParquetDataset/
├── Column1/
│   ├── file1.parquet
│   └── file2.parquet
└── Column2/
    ├── file1.parquet
    └── file2.parquet
```

Ex. columnar + row-wise chunking Parquet dataset.

# Tabular data · Parquet

- Could you:
  - Parse infinitely different groups of columns?
  - Scale better column-wise?

# Tabular data · Parquet

What else is inside Parquet? 👀

# Tabular data • Parquet

- **Definitions**:
  - **Schema**: "a representation of a plan or theory in the form of an outline or model." (Oxford Languages via Google)
  - **Metedata**: "data that provides information about other data" (Wikipedia: Metadata)

# Tabular data • Parquet

Okay, cool, what else? 🤔

- Data typing: every column has a type.
- Schema: can view types without inference
- Metadata: provide custom metadata

# Tabular data • Parquet

*Why are you bothering me about these things?!*

- Data typing: no second guessing a column type!
- Schema: documented data structure (see above)!
- Metadata: where/why did this data come to be?!

# Quick Technical Demonstration

A quick demonstration of these things in Parquet.

Google Colab Notebook

Jump back here when things get *awkward*. 🙃

# Tabular data · Dimensionality

- Dimensionality: Parquet supports a number of multidimensional data types which enhance our ability to share information about the data.

  Controversial: most things aren't *perfectly* 2D!

# Tabular data • Dimensionality

Dimensionality: most things aren't *perfectly* 2D!

There are hidden regularity expectations with data.

What if the data aren't regular?

# Tabular data • Dimensionality

What if the data aren't regular?

We spend ***a lot*** of time trying to make it regular. ☹️

# Tabular data • Dimensionality

- Concept: **Jagged arrays**
- "... a jagged array, also known as a ragged array or irregular array is an array of arrays of which the member arrays can be of different lengths ..." (Wikipedia: Jagged array)

# Tabular data • Dimensionality

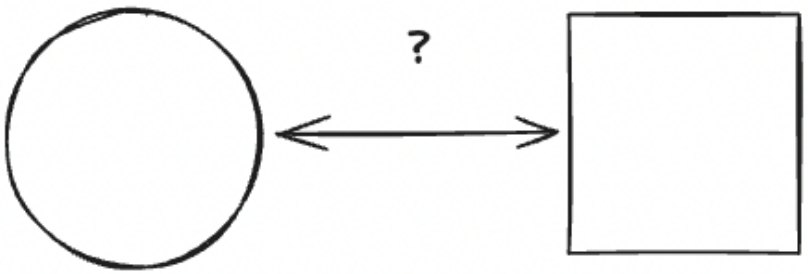
Jagged arrays in Python: Awkward Array.

# Tabular data • Dimensionality

Back to the Awkward demonstration.

Google Colab Notebook

# Tabular data • Dimensionality




Dimensionality can also be about **relationships**.

# Tabular data · Relationships

- To-do's:
  - Explore linked data and Parquet.
  - Investigate portable graph technologies for Parquet (Kuzu).
  - Expand understanding on query languages in context, SQL and Cypher.

**Thank you!**

Questions / Comments?