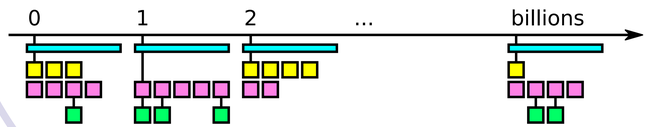
Awkward Arrays: Accelerating scientific data analysis on irregularly shaped data

Framework, #2103945 PI: Jim Pivarski

Awkward Array

An array library for **nested, variable-sized data**, including arbitrary-length lists, records, mixed types, and missing data, using **NumPy-like idioms**.



Arrays are **dynamically typed**, but operations on them are **compiled and fast**.

Coincides with NumPy when arrays are regular; generalizes when they're not.

What Awkward Array does

Example of an "awkward" array

```
array = ak.Array([
    [{"x": 1.1, "y": [1]}, {"x": 2.2, "y": [1, 2]}, {"x": 3.3, "y": [1, 2, 3]}],
    [],
    [{"x": 4.4, "y": [1, 2, 3, 4]}, {"x": 5.5, "y": [1, 2, 3, 4, 5]}]
])
```

Record structures with differently typed fields

Array of variable-length lists ("ragged" or "jagged" arrays)

Nested variable-length lists

Missing data

Heterogenous data (union/variant types)

NumPy-like expression

```
output = np.square(array["y", ..., 1:])
```

Result

```
output.to_list()
[
    [[], [4], [4, 9]],
    [],
    [[4, 9, 16], [4, 9, 16, 25]]
]
```

Equivalent Python

```
output = []
for sublist in python_objects:
    tmp1 = []
    for record in sublist:
        tmp2 = []
        for number in record["y"][1:]:
            tmp2.append(np.square(number))
        tmp1.append(tmp2)
    output.append(tmp1)
```

(for a similar calculation 10 million times larger, single threaded)

1.5 seconds
2.1 GB of memory

140 seconds
22 GB of memory

Library integrations

Awkward Array is part of a larger scientific Python ecosystem. It must work well with other libraries.

APACHE ARROW / Parquet

Lossless, bidirectional, zero-copy conversions between Awkward Array, Apache Arrow, Feather, and Parquet.

Integrated into AnnData for single-cell genomics.

JIT-compile Awkward Arrays in C++ with cppyy.

— or —

Build arrays in C++ using low-dependency header-only libraries.

Numba

Efficiently iterate over or build Awkward Arrays in code JIT-compiled by Numba, on CPUs and GPUs.

ROOT Data Analysis Framework

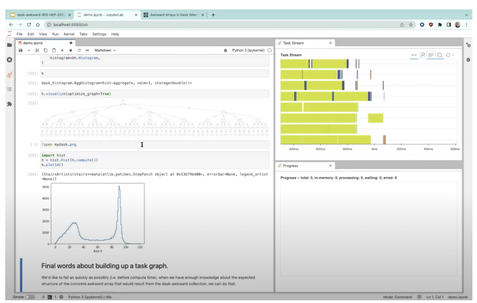
JIT-compiled as a C++ iterator in ROOT's RDataFrame workflow.

CuPy

Awkward Arrays on GPUs, backed by CuPy.

DASK

dask-awkward is a new high-level collection type for Dask, alongside Array and DataFrame.



JAX

Auto-differentiate through expressions involving Awkward Arrays, using JAX.

bluesky

Store Awkward Arrays in the Tiled database and slice them in the cloud.

pandas

Put an Awkward Array into a DataFrame column.

— or —

Explode an Awkward Array into a MultiIndex DataFrame.

Read any Kaitai Struct file format into Awkward Arrays.

Awkward Array.jl

Exchange data with the Julia language through a reimplementation of the Awkward Array memory layouts.

Adoption metrics

Awkward Array was developed in the context of particle physics; our current project is to find new applications in other fields of science and industry.

In addition to direct interactions with users, we scan GitHub for broad, statistical measures of adoption.

Words in user bios

Engineer
Scientist
Learning
Software
Data
Computer
University
Student
Research
Python
PhD
machine
enthusiast
Computing
Engineering
Developer
Working
Fellow
Postdoctoral
Candidate
Linux
Devops
NLP
physics
Optimization
Financial
Quantum
Intern
Intrest
Author
Fellow
Professor
Manager
Analyst
Tech
Work
Lover
Enthusiast
Lead
Senior
Scientific
Vision
Stuff
Week
Geek
Lab
Linux
Devops