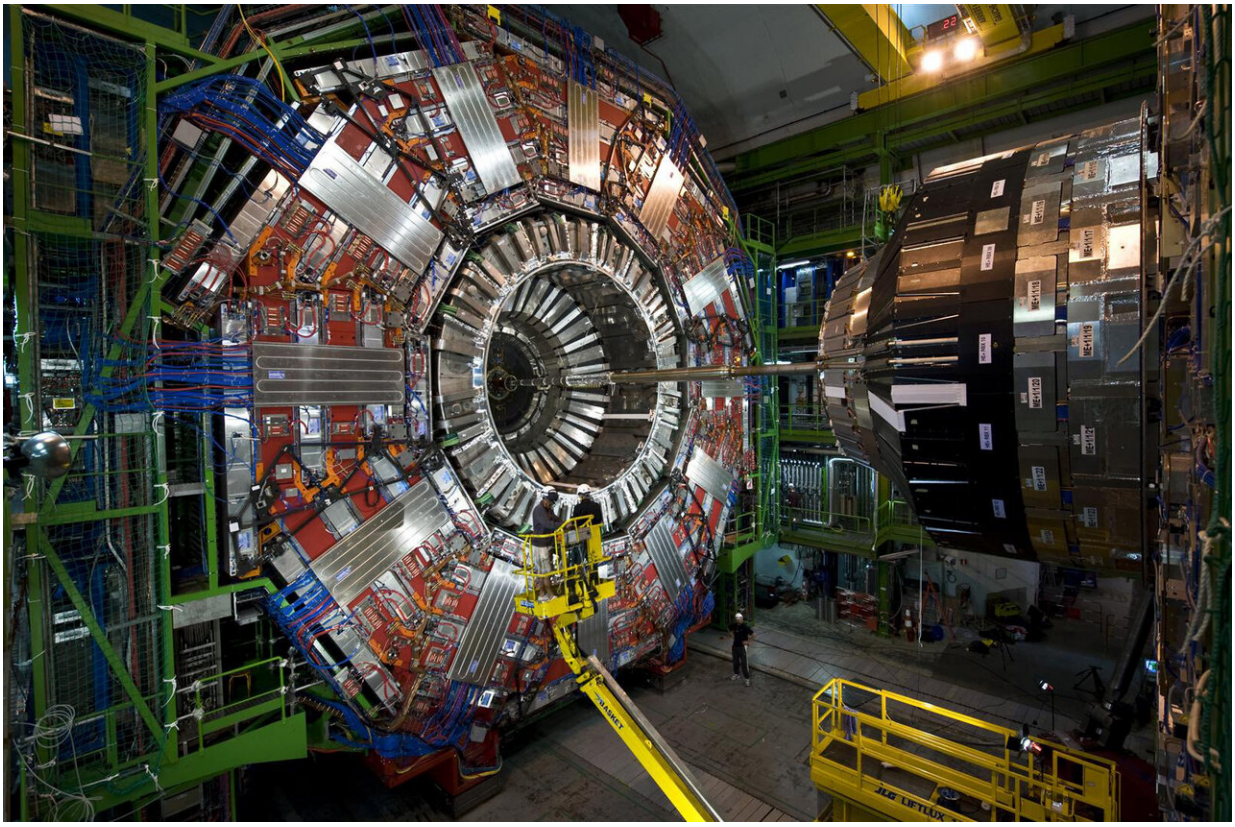
Coffea speeds up particle physics data analysis

February 22 2021



Around a dozen research groups on the CMS experiment at the Large Hadron Collider have adopted the Coffea data analysis tool for their work. Starting from information about the particles generated in collisions, Coffea enables large statistical analyses that hone researchers' understanding of the underlying physics, enabling faster run times and more efficient use of computing resources. Credit: CERN

Analyzing the mountains of data generated by the Large Hadron Collider at the European laboratory CERN takes so much time that even the computers need coffee. Or rather, Coffea—Columnar Object Framework for Effective Analysis.

A package in the programming language Python, Coffea (pronounced like the stimulating beverage) speeds up the analysis of massive data sets in high-energy physics research. Although Coffea streamlines computation, the software's primary goal is to optimize scientists' time.

"The efficiency of a human being in producing scientific results is of course affected by the tools that you have available," said Matteo Cremonesi, a postdoc at the U.S. Department of Energy's Fermi National Accelerator Laboratory. "If it takes more than a day for me to get a single number out of a computation—which often happens in high-energy physics—that's going to hamper my efficiency as a scientist."

Frustrated by the tedious manual work they faced when writing computer code to analyze LHC data, Cremonesi and Fermilab scientist Lindsey Gray assembled a team of Fermilab researchers in 2018 to adapt cutting-edge big data techniques to solve the most challenging questions in high-energy physics. Since then, around a dozen research groups on the CMS experiment—one of the LHC's two large general-purpose detectors—have adopted Coffea for their work.

Starting from information about the particles generated in collisions, Coffea enables large statistical analyses that hone researchers' understanding of the underlying physics. (Data processing facilities at the LHC carry out the initial conversion of raw data into a format particle physicists can use for analysis.) A typical analysis on the current LHC data set involves processing an astounding roughly 10 billion particle events that can add up to over 50 terabytes of data. That's the data equivalent of approximately 25,000 hours of streaming video on

Netflix.

At the heart of Fermilab's analysis tool lies a shift from a method known as event loop analysis to one called columnar analysis.

"You have a choice whether you want to iterate over each row and do an operation within the columns or if you want to iterate over the operations you're doing and attack all the rows at once," explained Fermilab postdoctoral researcher Nick Smith, the main developer of Coffea. "It's sort of an order-of-operations thing."

For example, imagine that for each row, you want to add together the numbers in three columns. In event loop analysis, you would start by adding together the three numbers in the first row. Then you would add together the three numbers in the second row, then move on to the third row, and so on. With a columnar approach, by contrast, you would start by adding the first and second columns for all the rows. Then you would add that result to the third column for all the rows.

"In both cases, the end result would be the same," Smith said. "But there are some trade-offs you make under the hood, in the machine, that have a big impact on efficiency."

In data sets with many rows, columnar analysis runs around 100 times faster than event loop analysis in Python. Yet prior to Coffea, particle physicists primarily used event loop analysis in their work—even for data sets with millions or billions of collisions.

The Fermilab researchers decided to pursue a columnar approach, but they faced a glaring challenge: High-energy physics data cannot easily be represented as a table with rows and columns. One particle collision might generate a slew of muons and few electrons, while the next might produce no muons and many electrons. Building on a library of Python

code called Awkward Array, the team devised a way to convert the irregular, nested structure of LHC data into tables compatible with columnar analysis. Generally, each row corresponds to one collision, and each column corresponds to a property of a particle created in the collision.

Coffea's benefits extend beyond faster run times—minutes compared to hours or days with respect to interpreted Python code—and more efficient use of computing resources. The software takes mundane coding decisions out of the hands of the scientists, allowing them to work on a more abstract level with fewer chances to make errors.

"Researchers are not here to be programmers," Smith said. "They're here to be data scientists."

Cremonesi, who searches for dark matter at CMS, was among the first researchers to use Coffea with no backup system. At first, he and the rest of the Fermilab team actively sought to persuade other groups to try the tool. Now, researchers frequently approach them asking how to apply Coffea to their own work.

Soon, Coffea's use will expand beyond CMS. Researchers at the Institute for Research and Innovation in Software for High Energy Physics, supported by the U.S. National Science Foundation, plan to incorporate Coffea into future analysis systems for both CMS and ATLAS, the LHC's other large general-purpose experimental detector. An upgrade to the LHC known as the High-Luminosity LHC, targeted for completion in the mid-2020s, will record about 100 times as much data, making the efficient data analysis offered by Coffea even more valuable for the LHC experiments' international collaborators.

In the future, the Fermilab team also plans to break Coffea into several Python packages, allowing researchers to use just the pieces relevant to

them. For instance, some scientists use Coffea mainly for its histogram feature, Gray said.

For the Fermilab researchers, the success of Coffea reflects a necessary shift in particle physicists' mindset.

"Historically, the way we do science focuses a lot on the hardware component of creating an experiment," Cremonesi said. "But we have reached an era in physics research where handling the software component of our scientific process is just as important."

Coffea promises to bring high-energy physics into sync with recent advances in big data in other scientific fields. This cross-pollination may prove to be Coffea's most far-reaching benefit.

"I think it's important for us as a community in high-energy physics to think about what kind of skills we're imparting to the people that we're training," Gray said. "Making sure that we as a field are pertinent to the rest of the world when it comes to data science is a good thing to do."

Provided by Fermi National Accelerator Laboratory