New data science platform speeds up Python queries

1 July 2021

Credit: CC0 Public Domain

Researchers from Brown University and MIT have developed a new data science framework that allows users to process data with the programming language Python—without paying the 'performance tax' normally associated with a user-friendly language.

The new framework, called Tuplex, is able to process data queries written in Python up to 90 times faster than industry-standard data systems like Apache Spark or Dask. The research team unveiled the system in research presented at SIGMOD 2021, a premier data processing conference, and have made the software freely available to all.

"Python is the primary programming language used by people doing data science," said Malte Schwarzkopf, an assistant professor of computer science at Brown and one of the developers of Tuplex. "That makes a lot of sense. Python is widely taught in universities, and it's an easy language to get started with. But when it comes to data science, there's a huge performance tax associated with Python because platforms can't process Python efficiently on the back end."

Platforms like Spark perform data analytics by distributing tasks across multiple processor cores or machines in a data center. That parallel processing allows users to deal with giant data sets that would choke a single computer to death. Users interact with these platforms by inputting their own queries, which contain custom logic written as "user-defined functions" or UDFs. UDFs specify custom logic, like extracting the number of bedrooms from the text of a real estate listing for a query that searches all of the real estate listings in the U.S. and selects all the ones with three bedrooms.

Because of its simplicity, Python is the language of choice for creating UDFs in the data science community. In fact, the Tuplex team cites a recent poll showing that 66% of data platform users utilize Python as their primary language. The problem is that analytics platforms have trouble dealing with those bits of Python code efficiently.

Data platforms are written in high-level computer languages that are compiled before running. Compilers are programs that take computer language and turn it into machine code—sets of instructions that a computer processor can quickly execute. Python, however, is not compiled beforehand. Instead, computers interpret Python code line by line while the program runs, which can mean far slower performance.

"These frameworks have to break out of their efficient execution of compiled code and jump into a Python interpreter to execute Python UDFs," Schwarzkopf said. "That process can be a factor of 100 less efficient than executing compiled code."

If Python code could be compiled, it would speed things up greatly. But researchers have tried for years to develop a general-purpose Python compiler, Schwarzkopf says, with little success. So instead of trying to make a general Python

compiler, the researchers designed Tuplex to compile a highly specialized program for the specific query and common-case input data. Uncommon input data, which account for only a small percentage of instances, are separated out and referred to an interpreter.

"We refer to this process as dual-case processing, as it splits that data into two cases," said Leonhard Spiegelberg, co-author of the research describing Tuplex. "This allows us to simplify the compilation problem as we only need to care about a single set of data types and common-case assumptions. This way, you get the best of two worlds: high productivity and fast execution speed."

And the runtime benefit can be substantial.

"We show in our research that a wait time of 10 minutes for an output can be reduced to a second," Schwarzkopf said. "So it really is a substantial improvement in performance."

In addition to speeding things up, Tuplex also has an innovative way of dealing with anomalous data, the researchers say. Large datasets are often messy, full of corrupted records or data fields that don't follow convention. In real estate data, for example, the number of bedrooms could either be a numeral or a spelled-out number. Inconsistencies like that can be enough to crash some data platforms. But Tuplex extracts those anomalies and sets them aside to avoid a crash. Once the program has run, the user then has the option of repairing those anomalies.

"We think this could have a major productivity impact for data scientists," Schwarzkopf said. "To not have to run out to get a cup of coffee while waiting for an output, and to not have a program run for an hour only to crash before it's done would be a really big deal."

More information: Paper: cs.brown.edu/people/malte/pub/ … 21-sigmod-tuplex.pdf

Software: tuplex.cs.brown.edu/

Provided by Brown University

APA citation: New data science platform speeds up Python queries (2021, July 1) retrieved 22 October 2021 from https://techxplore.com/news/2021-07-science-platform-python-queries.html