SDQL.py: An Efficient Query Engine in Python

Hesam Shahrokhi, Amir Shaikhha
School of Informatics, University of Edinburgh

Query Processing in Python: Current State of Affairs

Data Scientist

Pandas API
- Easy Coding
- Easy Debugging

Pandas Library
Low-Level Kernels
Prototyping

Hand-tuned Low-level Kernels
- No Global Optimization
- Slow Execution

Pandas/SDQL.py API
- Easy Coding
- Easy Debugging

SDQL.py Compiler
- Global Optimization
- Fast Execution

SQL in Python
- Hard to Code
- Hard to Debug

Physical Plan
Specialized Query Engine
Deployment

SDQL.py to SDQL IR Example

```python
@sdql_compile("R": R.type, "S": S.type), exec_mode)
def query(R, S):
    R.indx = R.sum(lambda r: (r.k: r.c) if r.a<9 else ())
    return S.sum(lambda s:
        (record("c": R.indx[s.k].c, "g": s.g)): True))
    if (s.e="pass" and R.indx[s.k]) else ()
```

Experimental Results

- Intel Core i5
- 16GB DDR4 RAM
- Ubuntu 20.04.3
- G++ 9.4.0, Python 3.8.10, and Pandas 1.4.1
- TPC-H Queries with Scaling Factor of 1