Accelerating Query Explanations Using Fine-Grained Provenance

Columbia University Data Science Institute

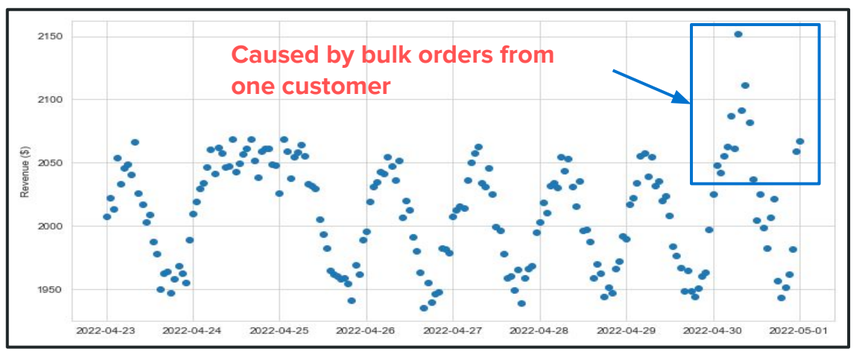
Alex Yao | Columbia University

Background & Motivation

Explanation engines empower users to ask “why” questions about their query results

Engines propose, execute, and evaluate **counterfactual interventions** to determine which most removes unexpected behavior in the output

```
SELECT time, sum(prices) FROM orders o JOIN customers c
WHERE c.name <> “John Doe”
GROUP BY time;
```



Problem Statement

Explanation engines are commonly bottlenecked by the evaluation of counterfactual interventions, which naively requires repeated executions of the original query. Existing approaches limit query and data complexity for performance.

Naive Approach 1: IVM

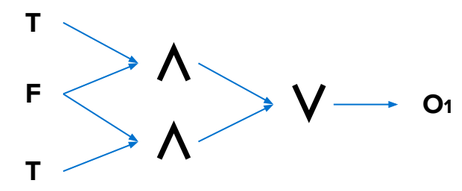
Long evaluation times for multi-row and multi-table interventions

$$(A - \nabla A) \bowtie (B - \nabla B) =$$

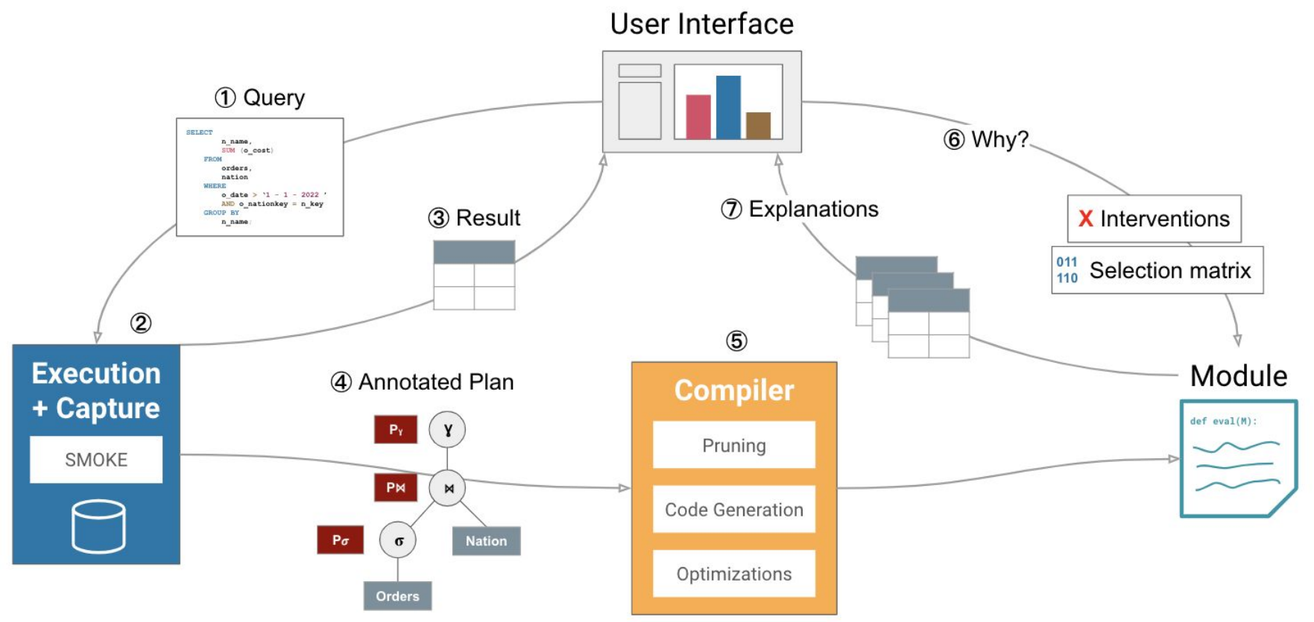
$$(A \bowtie B) - (\nabla A \bowtie B + A \bowtie \nabla B - \nabla A \bowtie \nabla B)$$

Naive Approach 2: Provenance Circuits

Conceptually great, but impractical, inefficient evaluation logic



Our Approach



*Our system uses **query operator provenance information** to generate evaluation code which quickly computes intervention results.*

LHS	RHS
1	2
1	4
3	3
…	…

Thread 1 → Join Exists? 1010

Thread 2 → 0111

Thread 3 → 0101

…

Code Generation

Input: LHS + RHS Existence Vectors

```
for i in range(# joined tuples):
    C[i] = A[LHS[i]] & B[RHS[i]]
```

Output: Join Output Existence Vector

Optimizations

- Vectorization
- Multithreading
- Provenance pruning

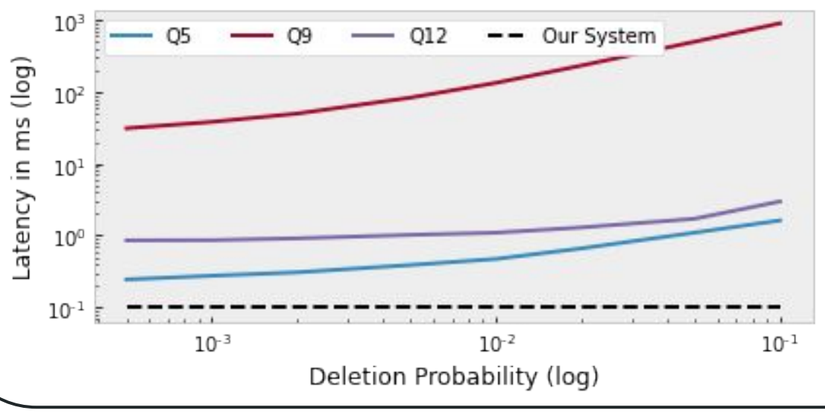
Experiments & Results

Dataset

TPC-H evaluated at SF 1, Interventions generated via randomized independent selections

vs. Circuits ProvSQL

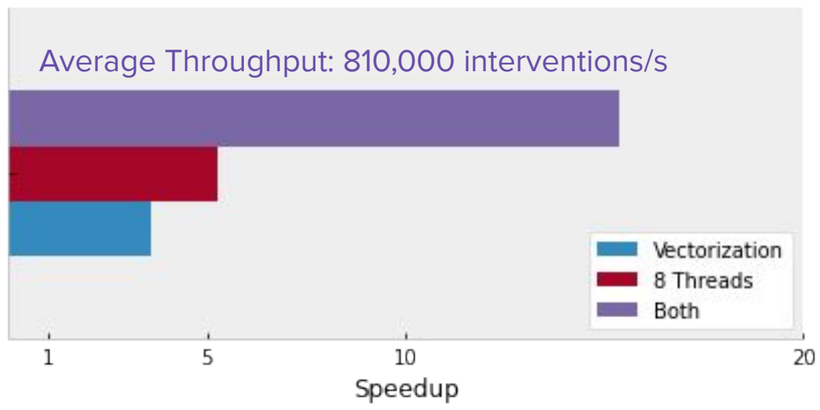
vs. IVM DBToaster*



Against best case IVM, we see significant wins over all cases, including simple aggregations for more than 10% deletions

* with our pruning applied

Performance in Perspective



Capable of handling over **1 million interventions a second** on analytical queries over multiple tables

Our system beats IVM and existing provenance systems by **600x** and **10,000x**