Eugene Wu

Assistant Professor

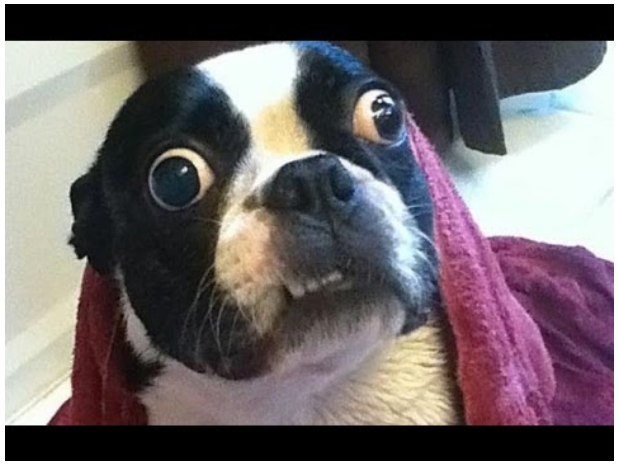
ewu@cs.columbia.edu

CUDBG

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Art

Legal Auditing



Showdog Pedigree
(stud book)

Provenance

What is it?

Projects@Columbia

Open Problems

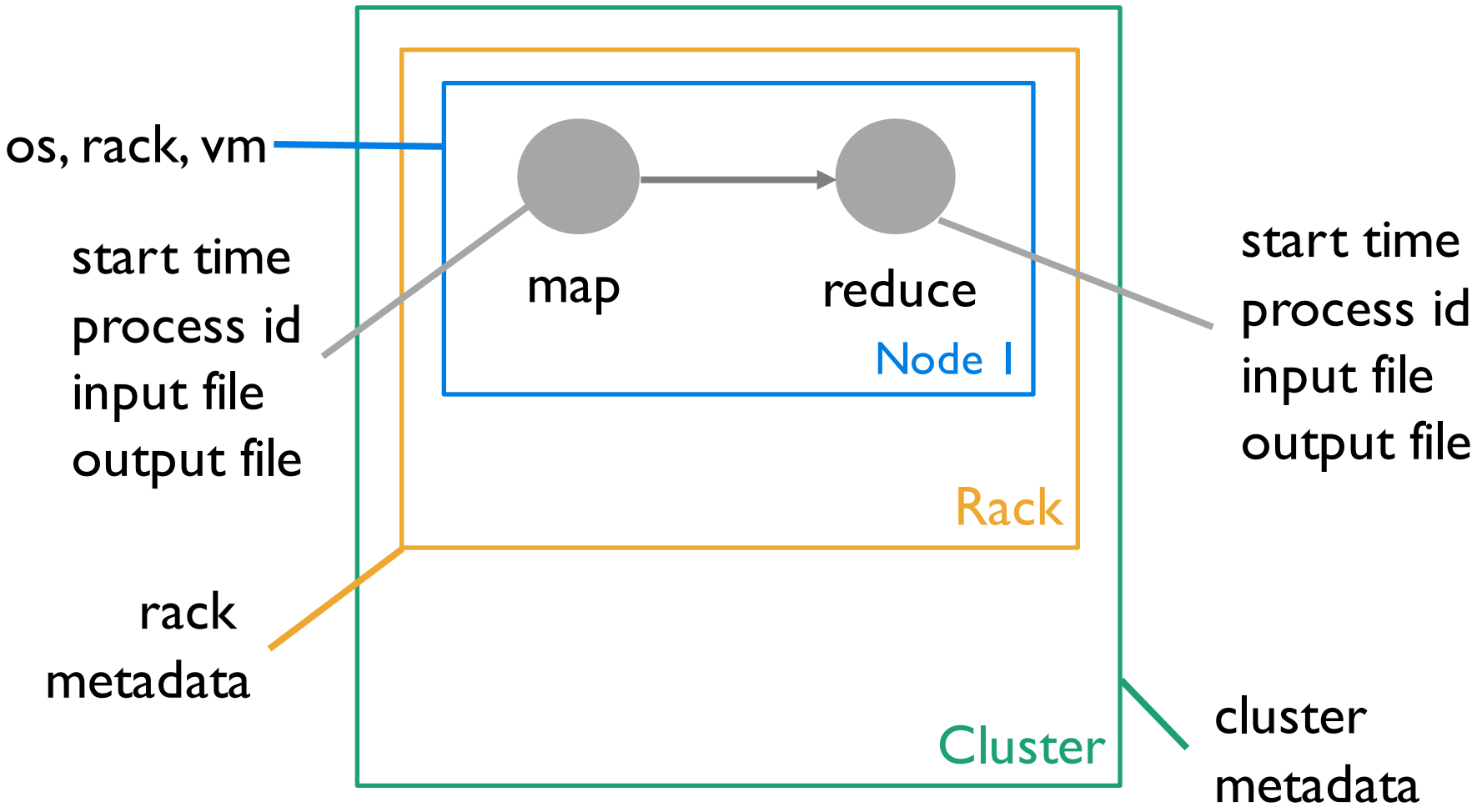
What Provenance Information is in a Map-Reduce job?

Static Information

map reduce

wordcount.java
file version

Dynamic Information



Answerable Queries

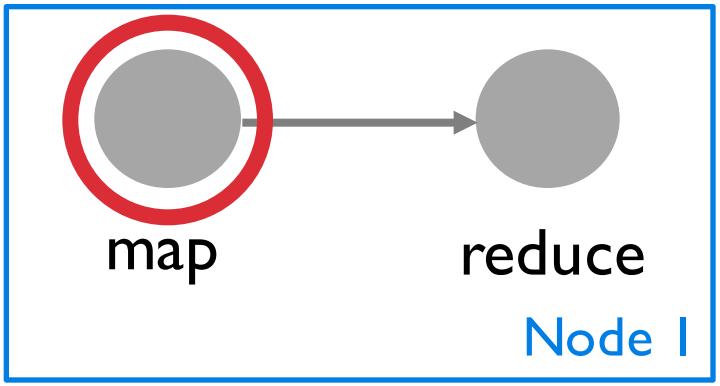
About Metadata

- What processes failed?
- What machines were they running on?
- What were inputs to these processes?
- What was their source code?

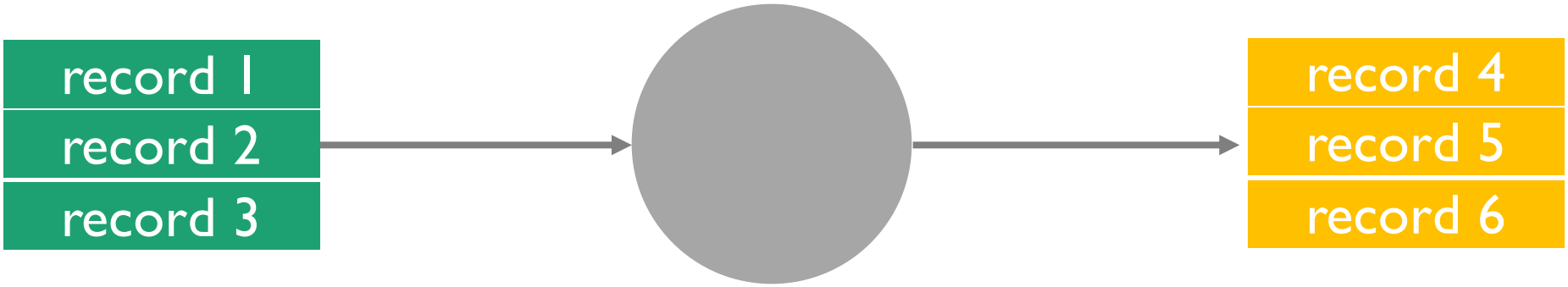
About Data

- This record is wrong, results are contaminated?
- This result val is wrong, what input records to check?

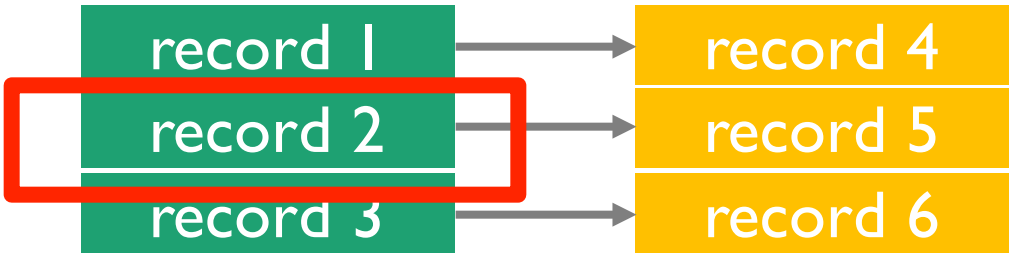
Lineage



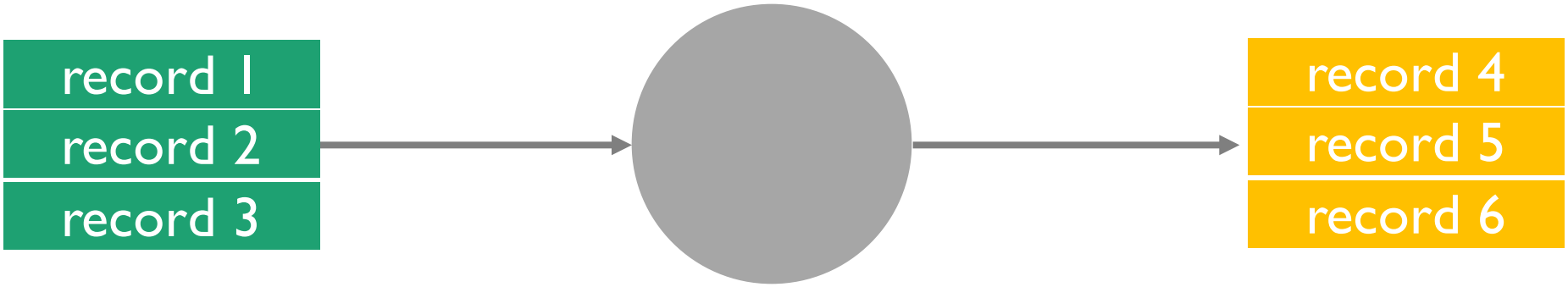
Lineage



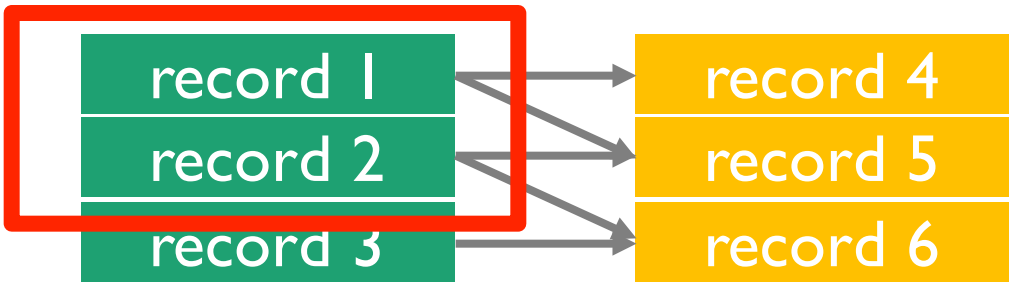
Which input records generated record 5 ?



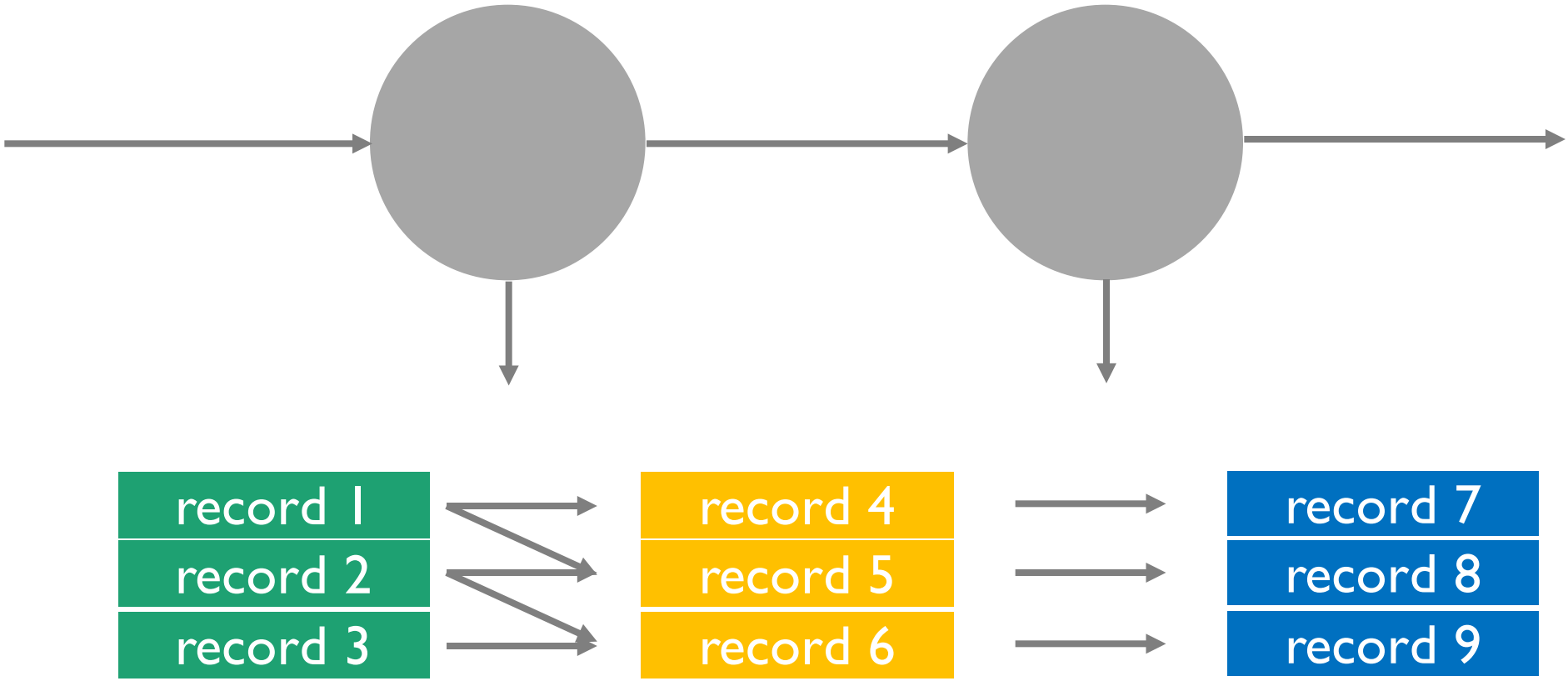
Lineage



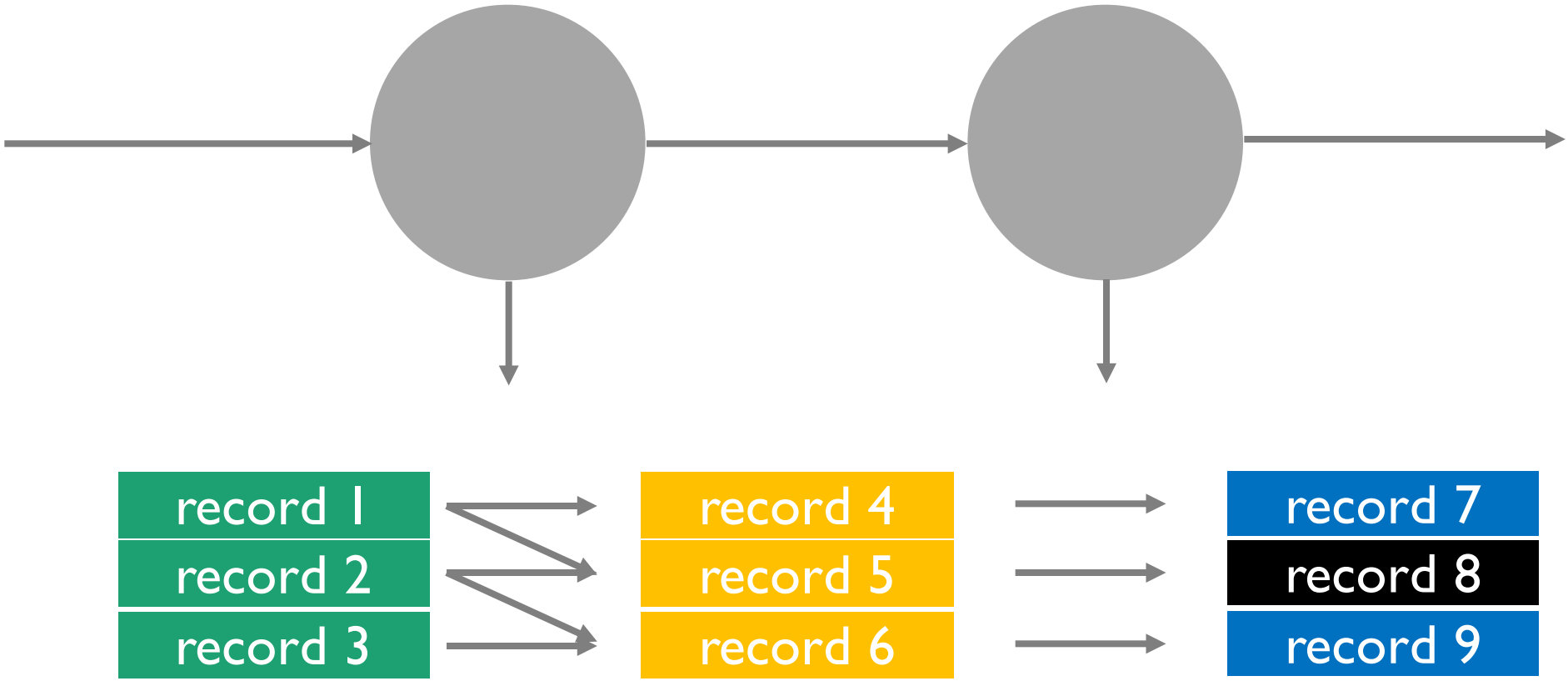
Which input records generated record 5 ?



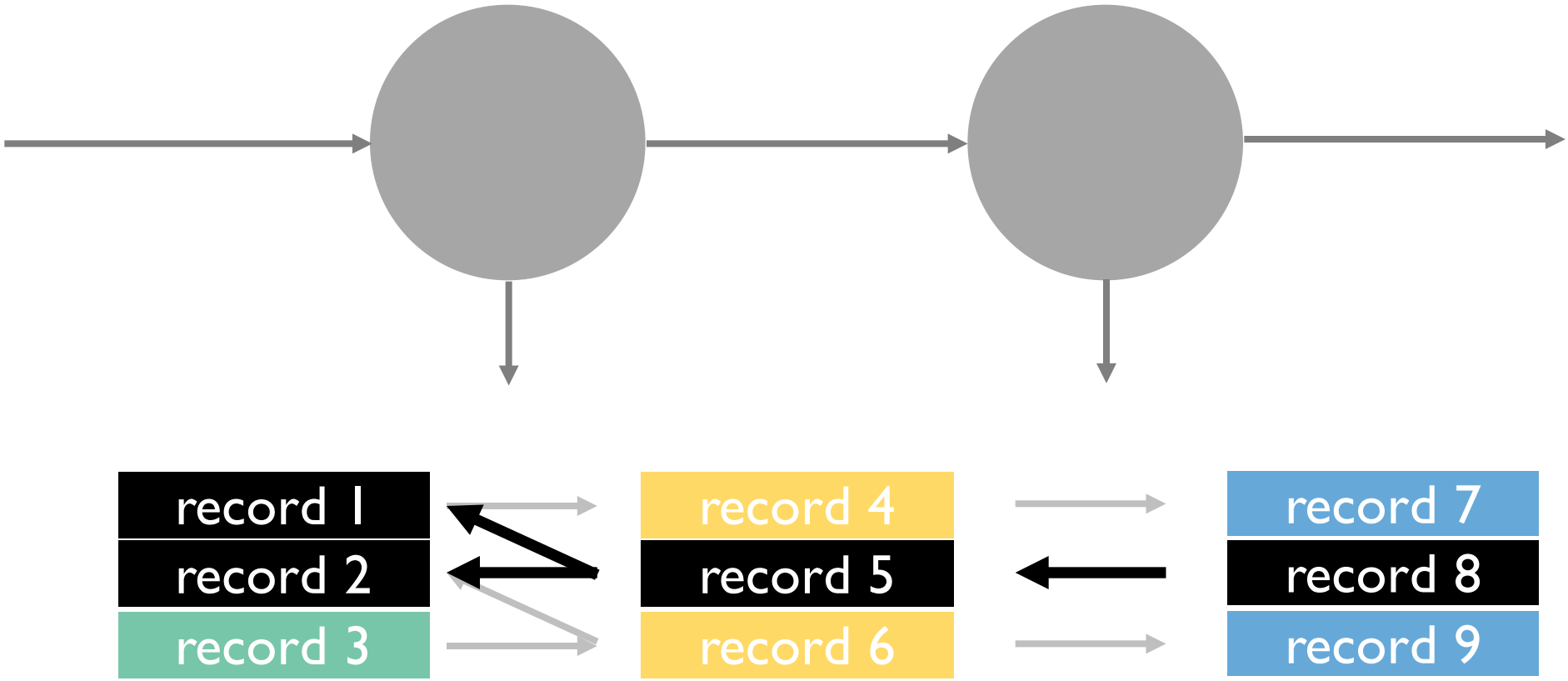
Lineage


record 1
record 2
record 3
record 4
record 5
record 6
record 7
record 8
record 9

Lineage


record 1
record 2
record 3
record 4
record 5
record 6
record 7
record 8
record 9

Lineage


record 1
record 2
record 3
record 4
record 5
record 6
record 7
record 8
record 9

```
sum = 0
for cell in input:
  if cell.type == 'valid':
    sum += cell.value
return sum
```

What cells were read to compute the result?

Return all cell

What values were added to compute the result?

Return all valid cells

Compute framework can

Provenance refers to all of this

Depends on program and lineage queries (hard)

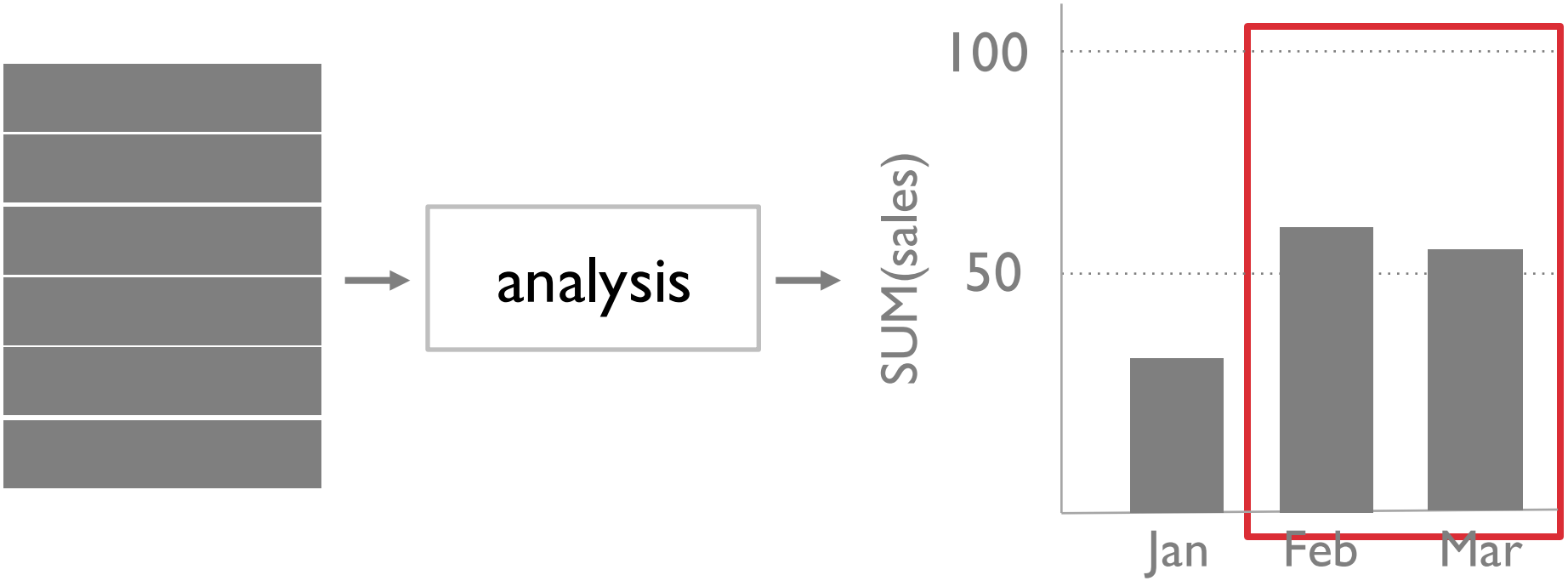
Provenance

What is it?

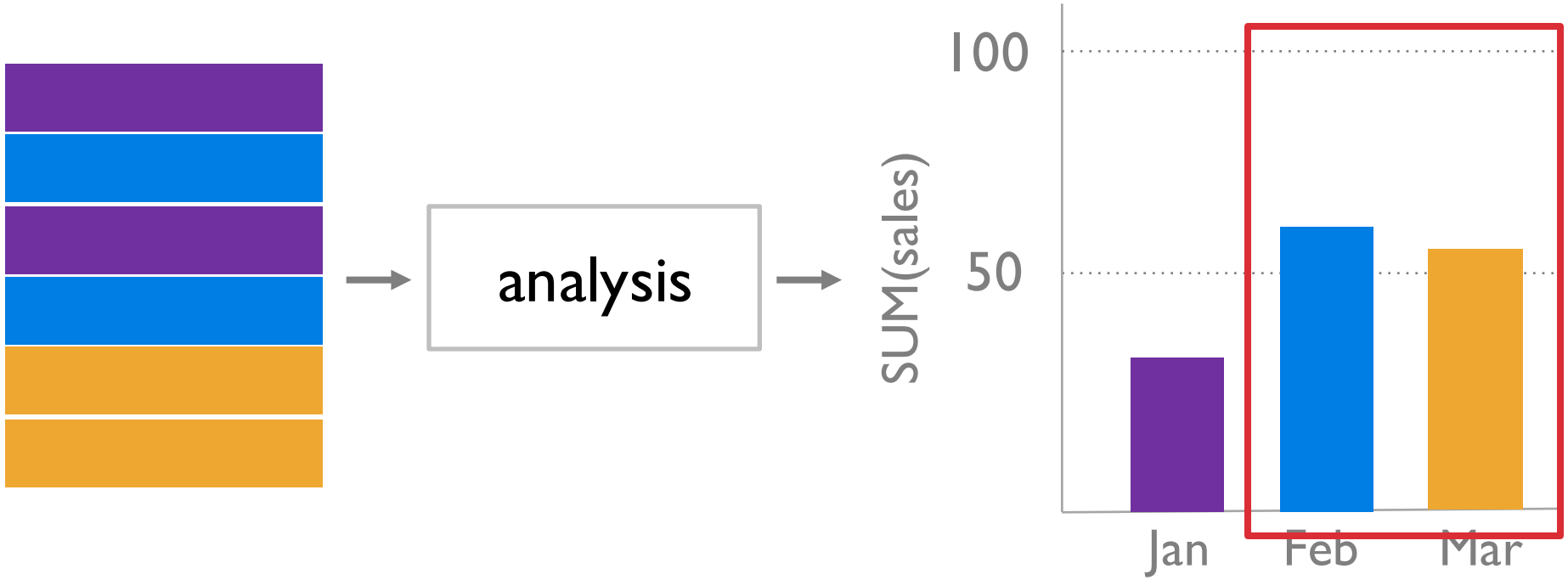
Projects@Columbia

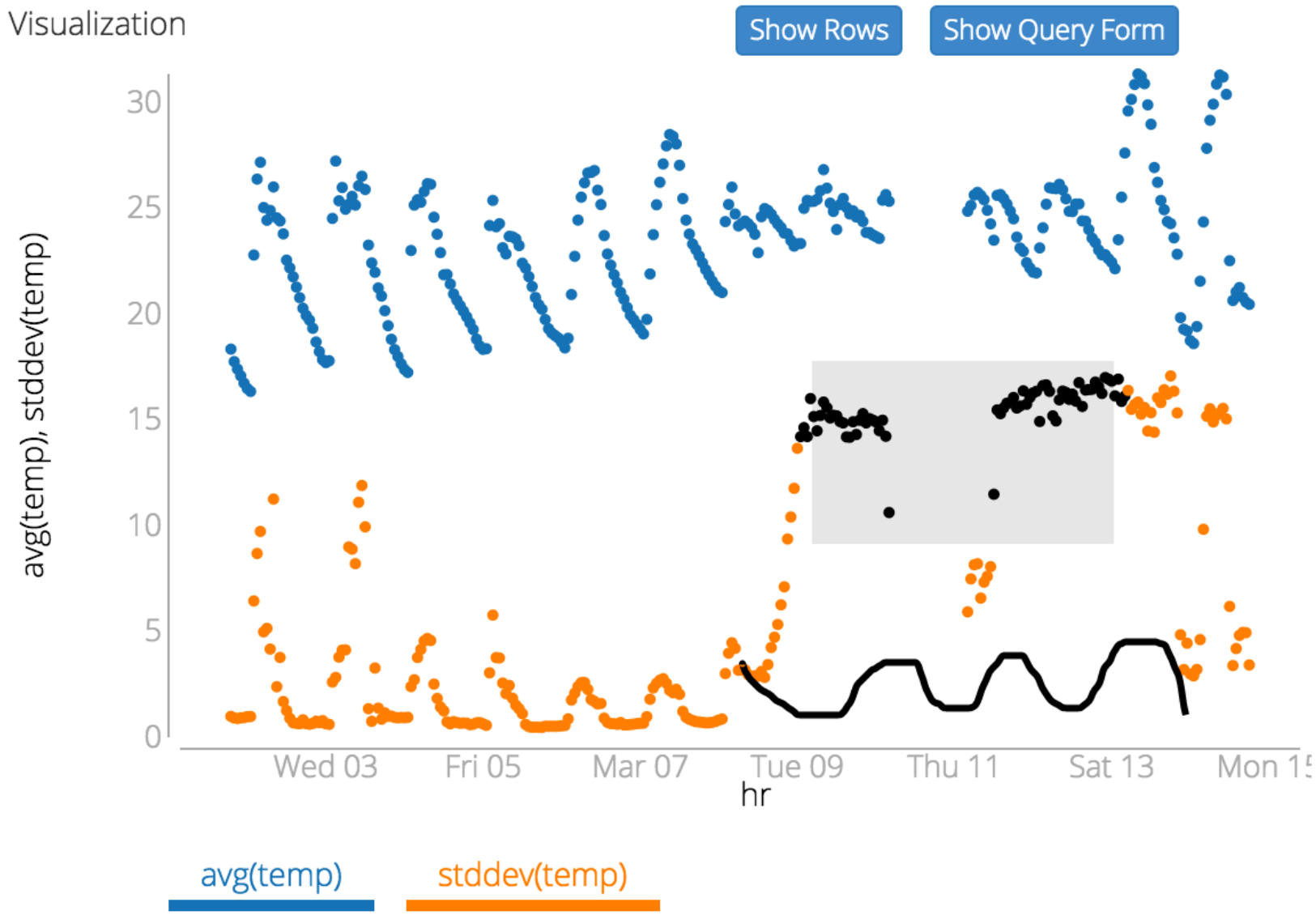
Open Problems

Scorpion: Explain Data



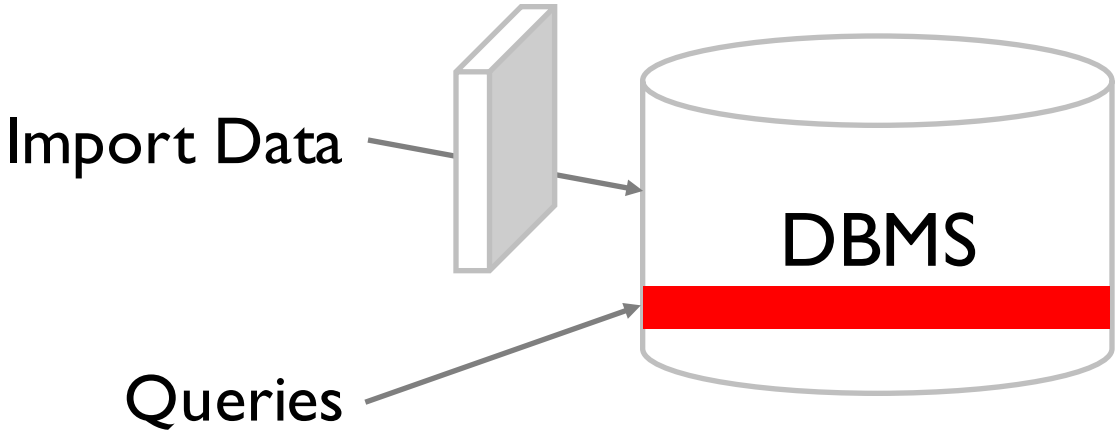
Scorpion: Explain Data

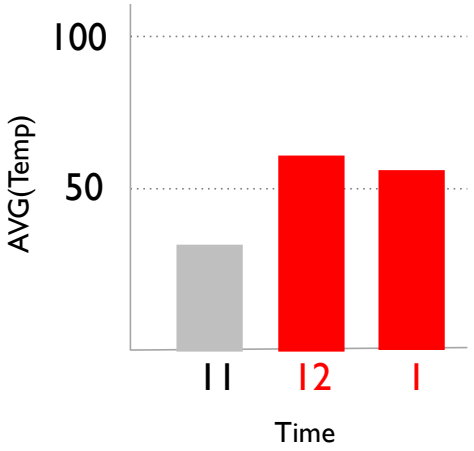


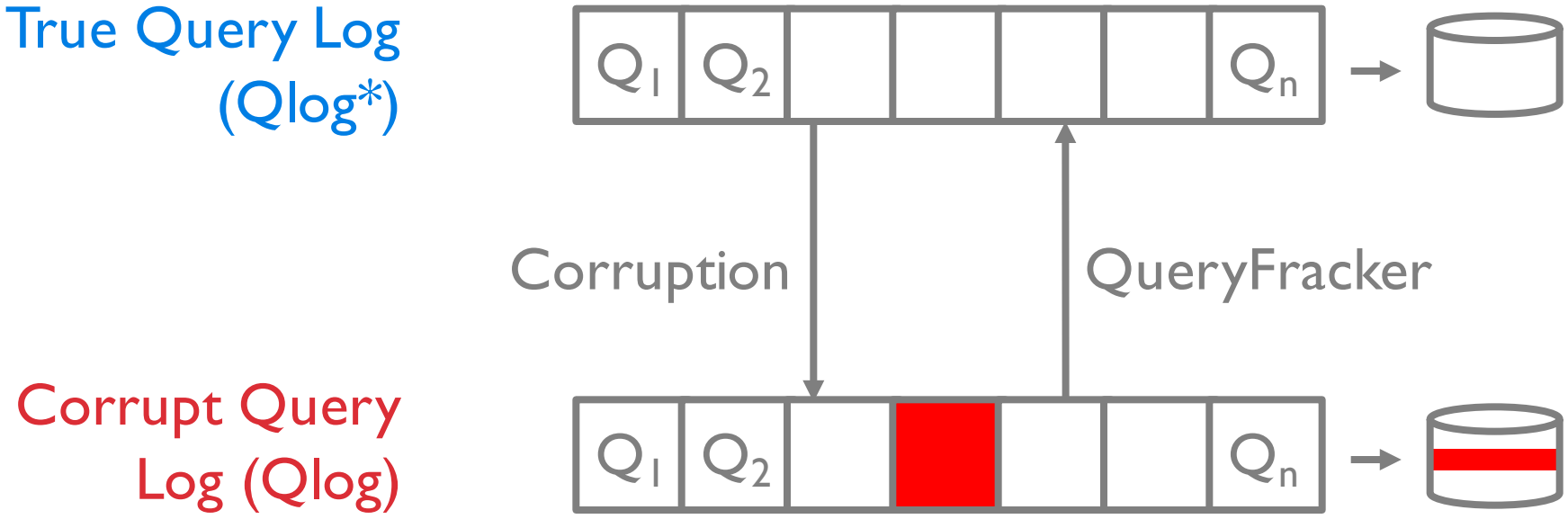


"Sensors near the window"

QueryFracker: Explain Queries

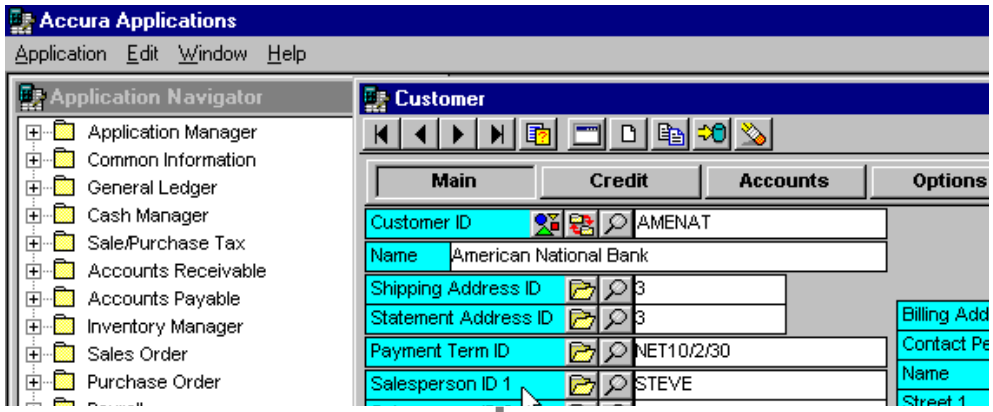




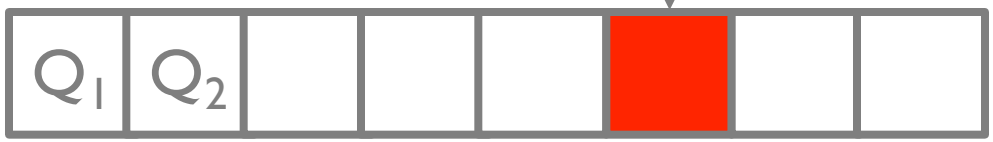


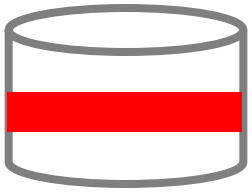
Goal

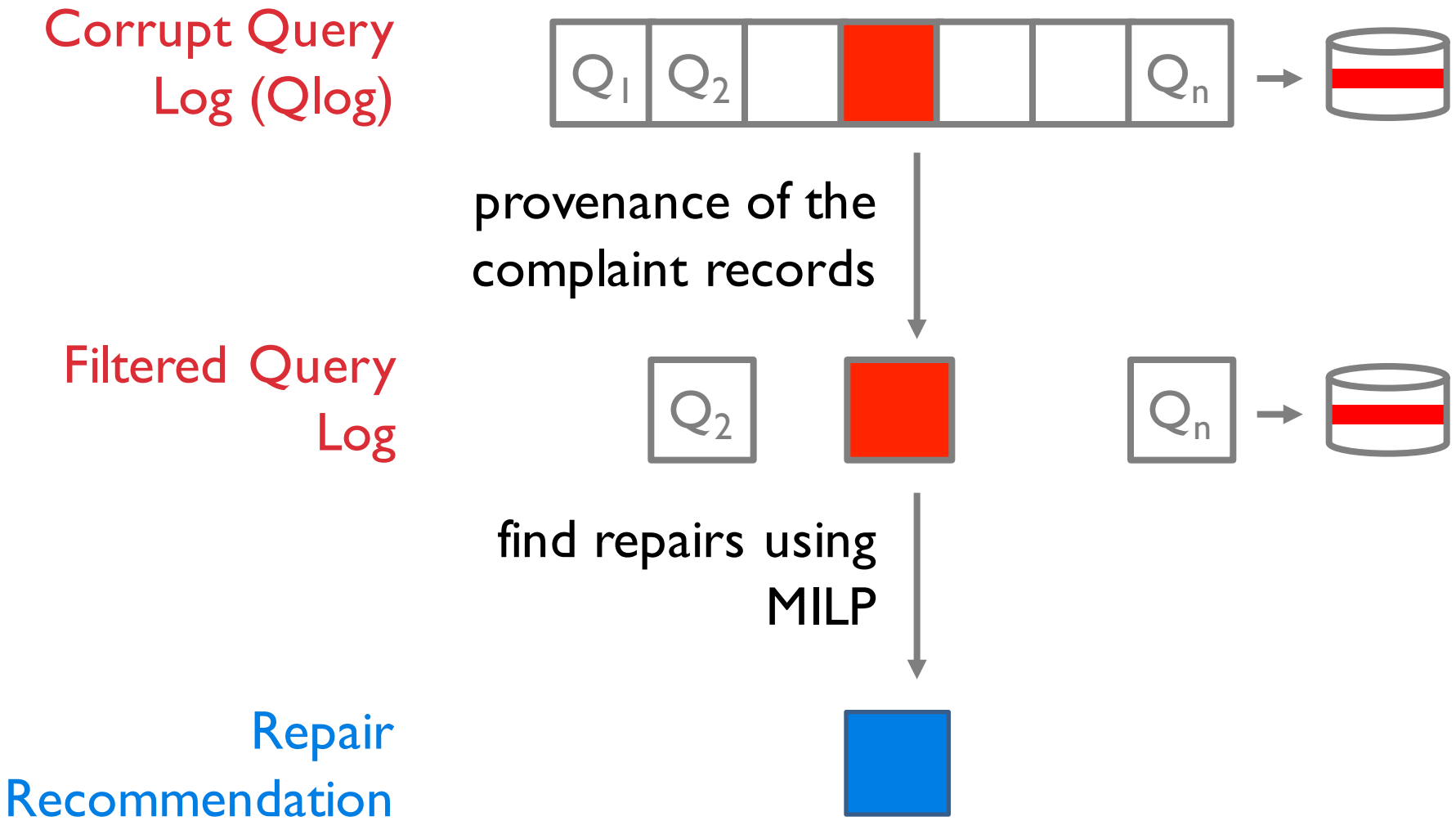
Given (Qlog, complaints) → recover Qlog*



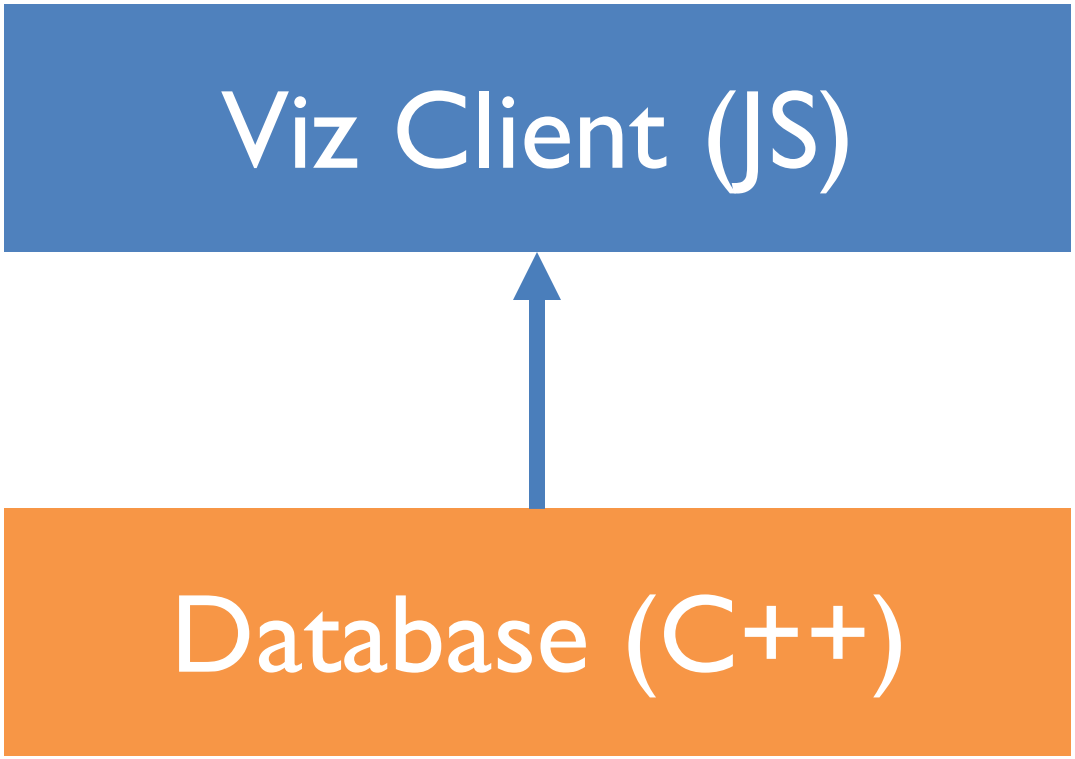
Application Server






Corrupt Query Log (Qlog)
Q1
Q2
Qn
provenance of the complaint records
Filtered Query Log
Q2
Qn
find repairs using MILP
Repair Recommendation

Data Visualization Management System



Data Visualization Management System

Analysis Tools

Provenance

Performance-aware Design

DEclarative Visual Interaction Language (DEVIL)

Design-aware Optimization

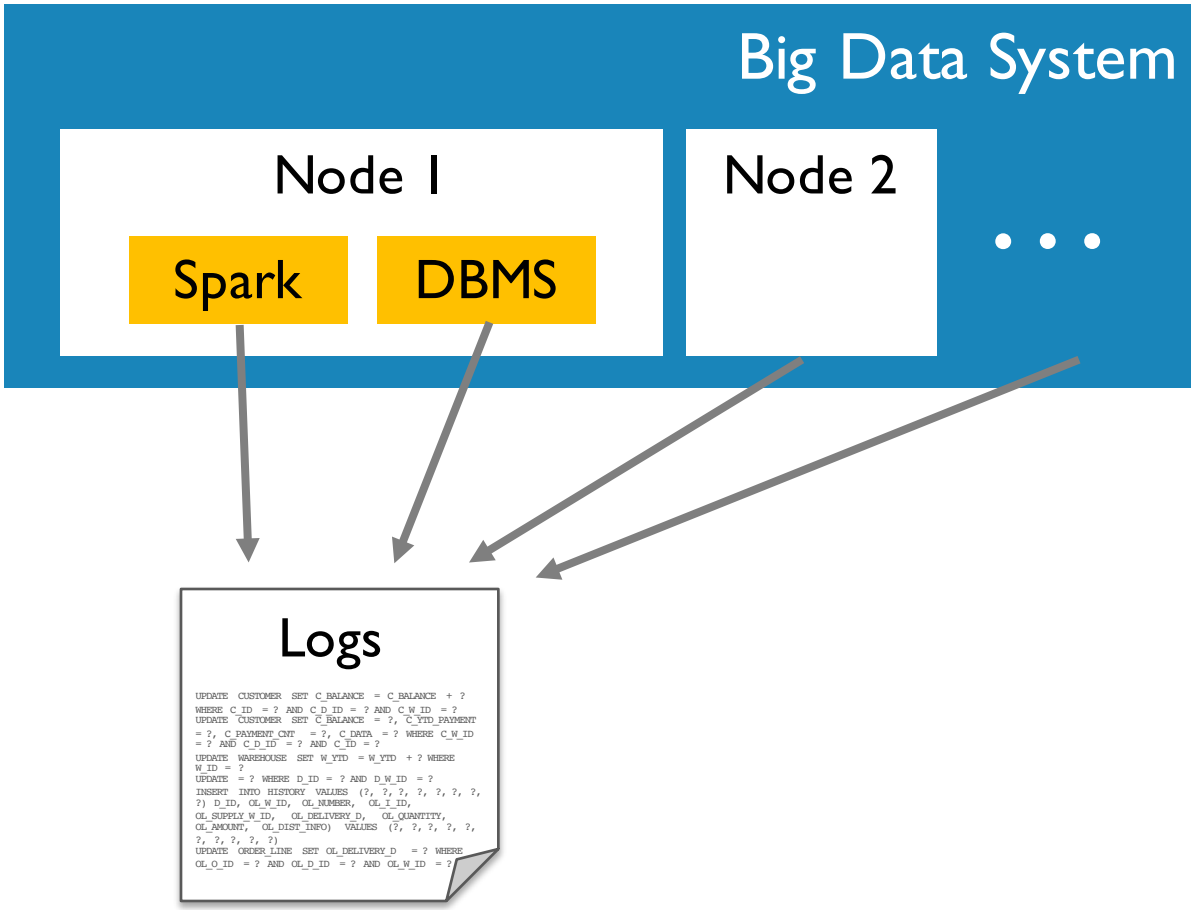
Provenance

What is it?

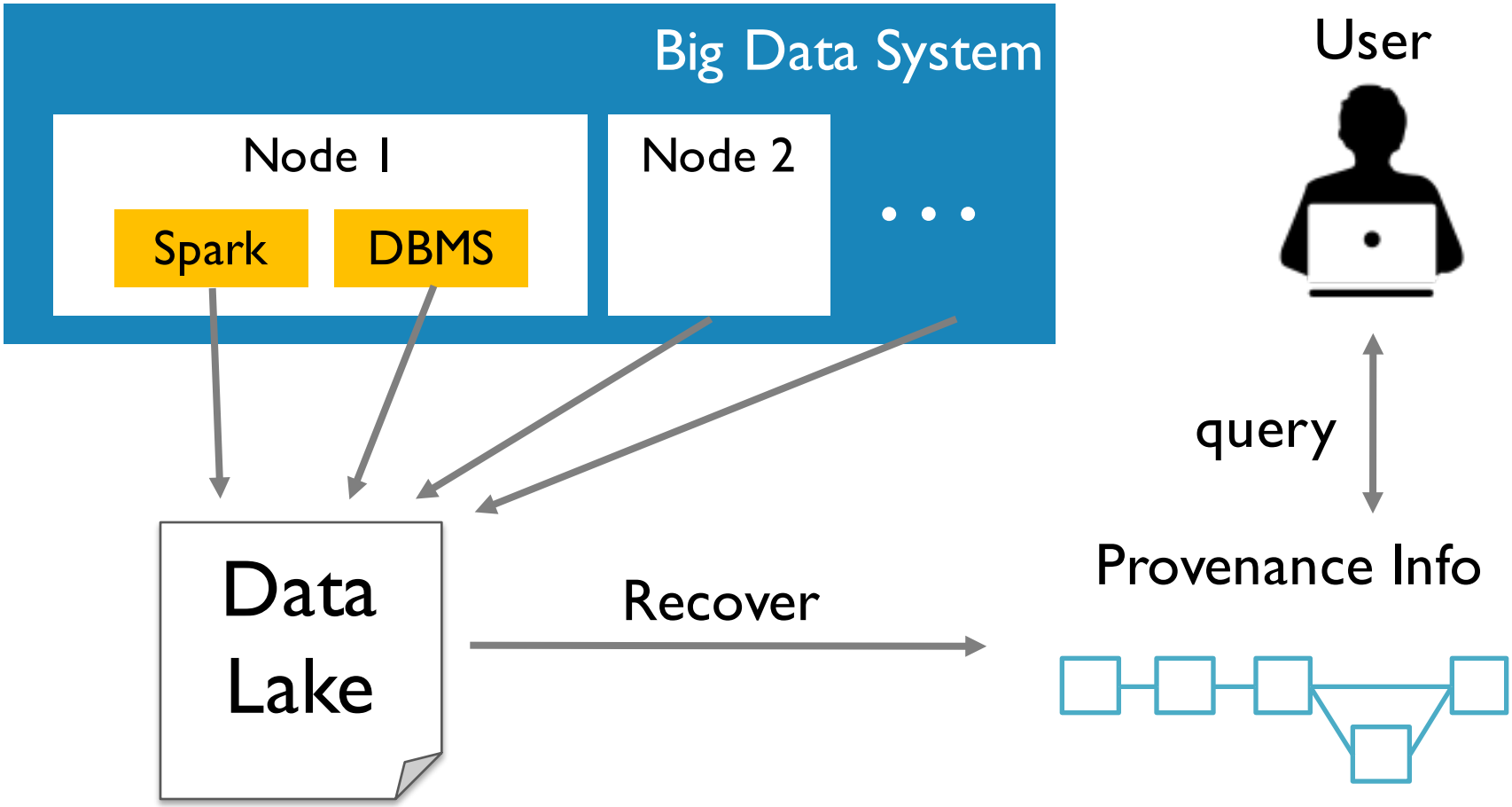
Projects@Columbia

Open Problems

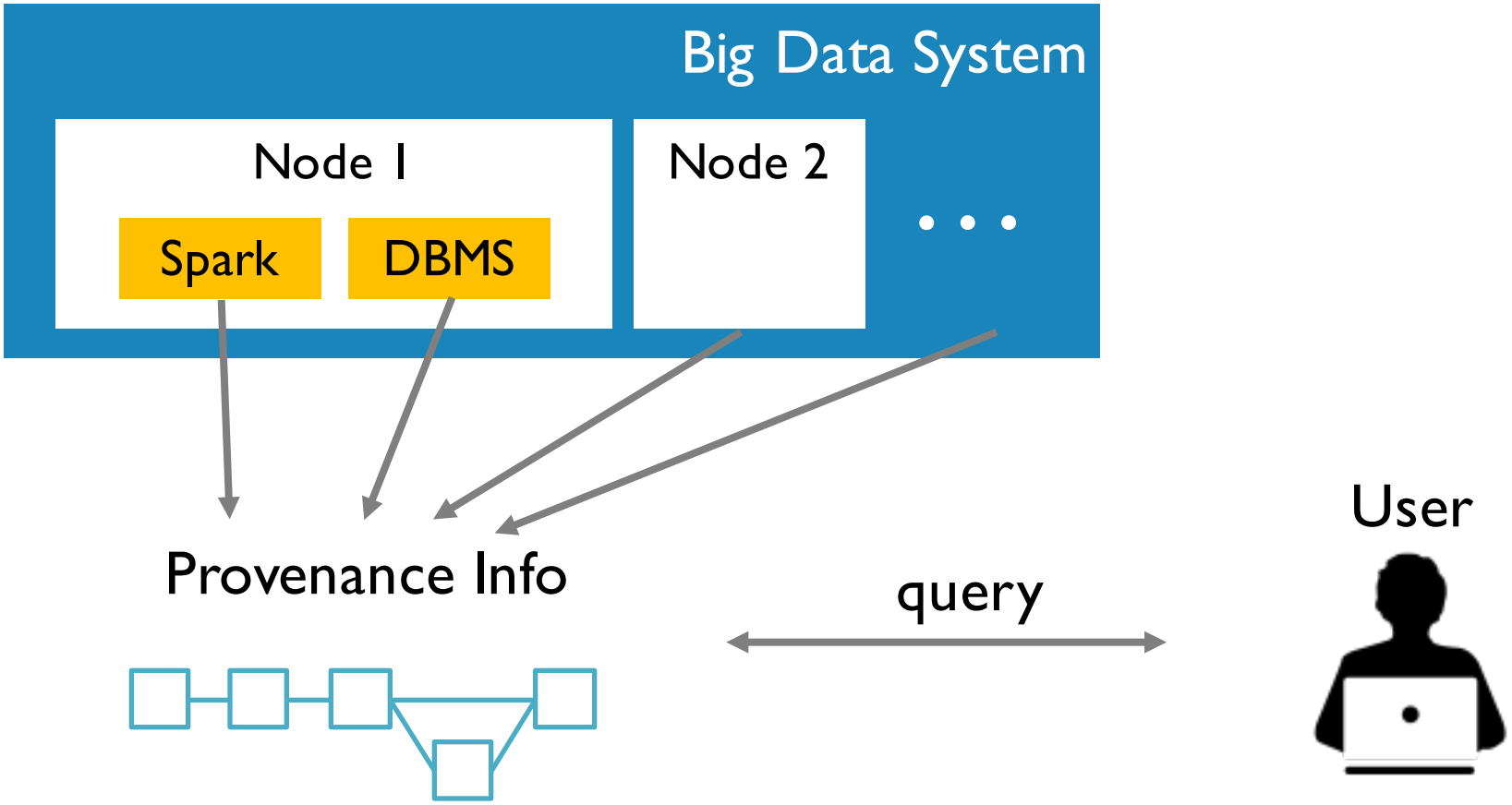
Recovering Provenance



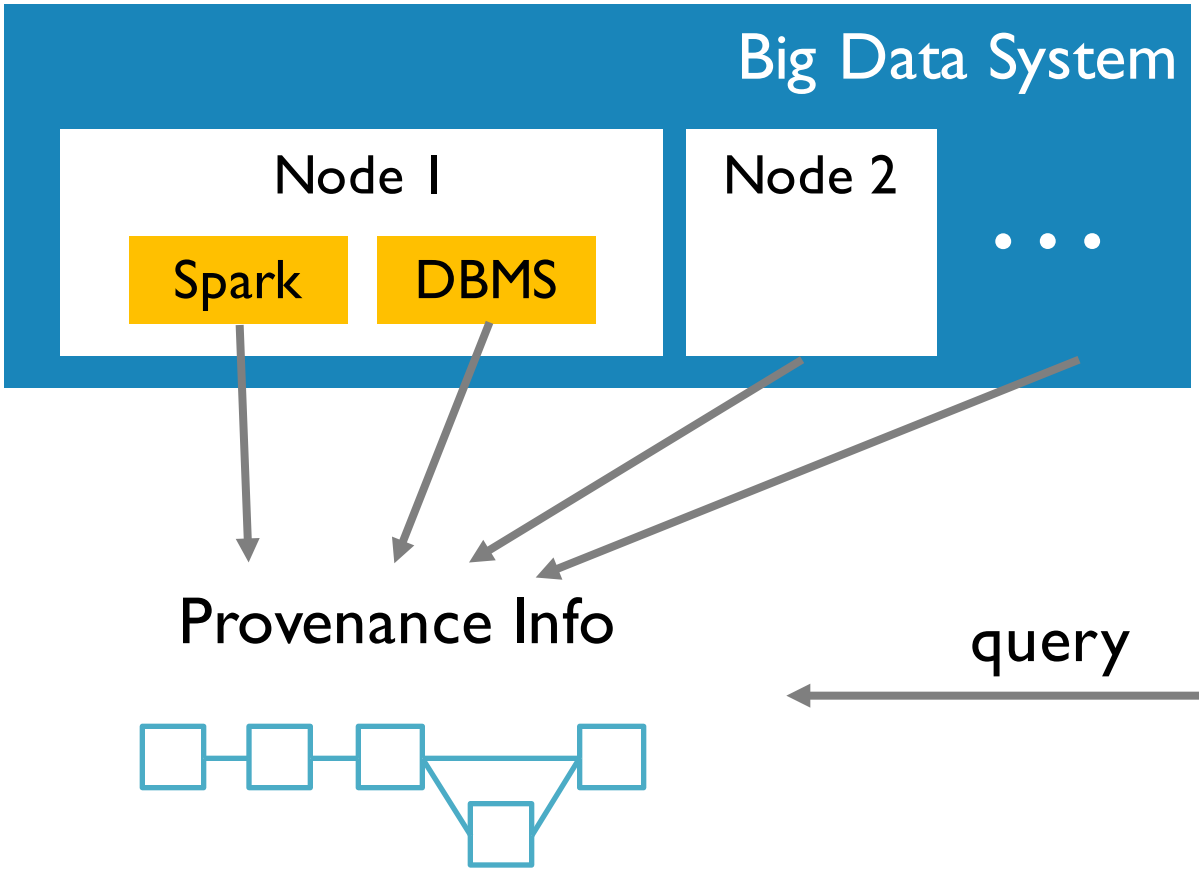
Recovering Provenance

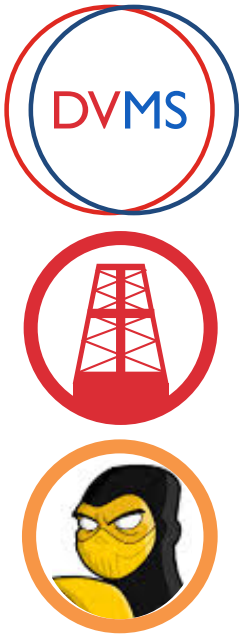


Query-based Instrumentation



Provenance-based Analysis





eugenewu.net

ewu@cs.columbia.edu

Explanation

Cleaning

Visualization

CUDBG

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK