SmokedDuck: Lineage via Selection Vector Capture

Charlie Summers | Haneen Mohammed | Eugene Wu
Columbia University

Introduction

Database Lineage can be used to improve a wide range of applications:
Query debugging, explanations, interactive visualizations, data cleaning, governance, etc.

Existing lineage capture techniques[1,2,3] are either too slow, or cannot be used with vectorized databases that are designed for big data, analytic workloads.

What is Lineage?

The relationship between the input rows and output rows of a query.

```
SELECT GradYear, AVG(GPA)
FROM Students
WHERE GPA > 3.0
GROUP BY GradYear
```

Students

Name	GradYear	GPA
Sue	2023	3.2
Beau	2024	2.7
Jungmin	2023	3.6
Harish	2023	3.8
Scott	2024	2.4
Eliza	2024	3.4

WHERE GPA > 3.0

Name	GradYear	GPA
Sue	2023	3.2
Jungmin	2023	3.6
Harish	2023	3.8
Eliza	2024	3.4

GROUP BY GradYear

GradYear	GPA
2023	3.53
2024	3.4

Selection Vector Capture

GPA
3.2
2.7
3.6
3.8
2.4
3.4

GPA > 3.0

SelVec
1
3
4
6

GradYear
2023
2024
2023
2023
2024
2024

GradYear'
2023
2023
2023
2024

The Selection Vector encodes the Lineage for this operator.

Lineage Use Case: Explanations

```
SELECT GradYear, AVG(GPA)
FROM Students
GROUP BY GradYear
```

Result

GradYear	GPA
2023	3.53
2024	2.83

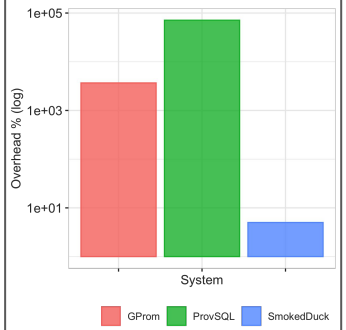
Students

Name	GradYear	GPA
Sue	2023	3.2
Beau	2024	2.7
Jungmin	2023	3.6
Harish	2023	3.8
Scott	2024	2.4
Eliza	2024	3.4

```
Why is GPA for
GradYear 2024 so low?
```

```
SELECT *
FROM Students
WHERE GPA < 2.8
```

Lineage Capture Speed



Lineage Capture on Group By query with 10M rows

[1] B. S. Arab, S. Feng, B. Glavic, S. Lee, X. Niu, and Q. Zeng. Gprom - a swiss army knife for your provenance needs. A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering, 41(1), 2018.
[2] F. Psallidas and E. Wu. Smoke: Fine-grained lineage at interactive speed. PVLDB,11:719 - 732, 2018.
[3] P. Senellart, L. Jachiet, S. Maniu, and Y. Ramusat. Provsql: Provenance and probability management in postgresql. Proceedings of the VLDB Endowment(PVLDB), 11(12):2034–2037, 2018.