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 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.

Selection Vector Capture

GPA > 3.0

GradYear

SelVec

1
3
4
6

The Selection Vector encodes the Lineage for this operator.

Lineage Use Case: Explanations

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

<table>
<thead>
<tr>
<th>Name</th>
<th>GradYear</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sue</td>
<td>2023</td>
<td>3.2</td>
</tr>
<tr>
<td>Beau</td>
<td>2024</td>
<td>2.7</td>
</tr>
<tr>
<td>Jungmin</td>
<td>2023</td>
<td>3.6</td>
</tr>
<tr>
<td>Harish</td>
<td>2023</td>
<td>3.8</td>
</tr>
<tr>
<td>Scott</td>
<td>2024</td>
<td>2.4</td>
</tr>
<tr>
<td>Eliza</td>
<td>2024</td>
<td>3.4</td>
</tr>
</tbody>
</table>

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