**Opaque:** An Oblivious and Encrypted Distributed Analytics Platform

[NSDI’17]

Wenting Zheng, Ankur Dave, Jethro Beekman, Raluca Ada Popa, Joseph Gonzalez, and Ion Stoica

UC Berkeley
Opaque*: oblivious and encrypted distributed analytics platform

* Oblivious Platform for Analytic QUEries
Opaque*: oblivious and encrypted distributed analytics platform

Leverages hardware enclaves (Intel SGX)

* Oblivious Platform for Analytic QUERies
Access patterns leakage
Access patterns

machine 0

processor → memory
Access patterns

machine 0

processor

addresses

memory
Access patterns

machine 0

processor

addresses

memory

network messages

machine 1
Example: network access pattern leakage
Example: network access pattern leakage

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
</tr>
</tbody>
</table>
Example: network access pattern leakage

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
</tr>
</tbody>
</table>

```
SELECT count(*) FROM medical
GROUP BY disease
```
Example: network access pattern leakage

<table>
<thead>
<tr>
<th></th>
<th>...</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td></td>
<td><strong>Diabetes</strong></td>
</tr>
<tr>
<td>29489</td>
<td></td>
<td><strong>Diabetes</strong></td>
</tr>
<tr>
<td>13744</td>
<td></td>
<td><strong>Cancer</strong></td>
</tr>
<tr>
<td>18740</td>
<td></td>
<td><strong>Diabetes</strong></td>
</tr>
<tr>
<td>98329</td>
<td></td>
<td><strong>Cancer</strong></td>
</tr>
<tr>
<td>32591</td>
<td></td>
<td><strong>Diabetes</strong></td>
</tr>
</tbody>
</table>
Example: network access pattern leakage

12809 ... Diabetes
29489 ... Diabetes
13744 ... Cancer

18740 ... Diabetes
98329 ... Cancer
32591 ... Diabetes
Example: network access pattern leakage

| 12809 | Diabetes |
| 29489 | Diabetes |
| 13744 | Cancer   |
| 18740 | Diabetes |
| 98329 | Cancer   |
| 32591 | Diabetes |
Example: network access pattern leakage

<table>
<thead>
<tr>
<th></th>
<th>Diabetes</th>
<th>Diabetes</th>
<th>Cancer</th>
<th>Cancer</th>
<th>Diabetes</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29489</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13744</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18740</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>98329</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32591</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example: network access pattern leakage

<table>
<thead>
<tr>
<th>12809</th>
<th>...</th>
<th>Diabetes</th>
</tr>
</thead>
<tbody>
<tr>
<td>29489</td>
<td>...</td>
<td>Diabetes</td>
</tr>
<tr>
<td>13744</td>
<td>...</td>
<td>Cancer</td>
</tr>
<tr>
<td>18740</td>
<td>...</td>
<td>Diabetes</td>
</tr>
<tr>
<td>98329</td>
<td>...</td>
<td>Cancer</td>
</tr>
<tr>
<td>32591</td>
<td>...</td>
<td>Diabetes</td>
</tr>
</tbody>
</table>
Example: network access pattern leakage
Example: network access pattern leakage

<table>
<thead>
<tr>
<th>12809</th>
<th>Diabetes</th>
</tr>
</thead>
<tbody>
<tr>
<td>129489</td>
<td>Diabetes</td>
</tr>
<tr>
<td>18740</td>
<td>Diabetes</td>
</tr>
<tr>
<td>32591</td>
<td>Diabetes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>13744</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>98329</td>
<td>Cancer</td>
</tr>
</tbody>
</table>

Public information:
Diabetes twice as common as cancer
Example: network access pattern leakage

Public information:
Diabetes twice as common as cancer
Example: network access pattern leakage

Public information: Diabetes twice as common as cancer
Example: network access pattern leakage

Public information:
Diabetes twice as common as cancer
Example: network access pattern leakage

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Diabetes</th>
<th>Cancer</th>
<th></th>
<th></th>
<th></th>
<th>Diabetes</th>
<th>Cancer</th>
<th></th>
<th></th>
<th></th>
<th>Diabetes</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>...</td>
<td></td>
<td>Diabetes</td>
<td>12809</td>
<td>...</td>
<td>Diabetes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29489</td>
<td>...</td>
<td>Diabetes</td>
<td></td>
<td>29489</td>
<td>...</td>
<td>Diabetes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13744</td>
<td>...</td>
<td>Cancer</td>
<td></td>
<td>18740</td>
<td>...</td>
<td>Diabetes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18740</td>
<td>...</td>
<td>Diabetes</td>
<td></td>
<td>13744</td>
<td>...</td>
<td>Cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>98329</td>
<td>...</td>
<td>Cancer</td>
<td></td>
<td>96329</td>
<td></td>
<td>Cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32591</td>
<td>...</td>
<td>Diabetes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example: network access pattern leakage
Example: network access pattern leakage

- 12809 ... Diabetes
- 29489 ... Diabetes
- 13744 ... Cancer

- 18740 ... Diabetes
- 98329 ... Cancer
- 32591 ... Diabetes
Example: network access pattern leakage

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Diabetes</th>
<th></th>
<th></th>
<th>Diabetes</th>
<th></th>
<th></th>
<th>Diabetes</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>...</td>
<td></td>
<td>29489</td>
<td>...</td>
<td></td>
<td>13744</td>
<td>...</td>
<td>Cancer</td>
</tr>
<tr>
<td>18740</td>
<td>...</td>
<td>Diabetes</td>
<td>32591</td>
<td>...</td>
<td>Diabetes</td>
<td>98329</td>
<td>...</td>
<td>Cancer</td>
</tr>
</tbody>
</table>
Example: network access pattern leakage

<table>
<thead>
<tr>
<th></th>
<th>Diabetes</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29489</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13744</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18740</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32591</td>
<td></td>
<td></td>
</tr>
<tr>
<td>98329</td>
<td></td>
<td>Cancer</td>
</tr>
</tbody>
</table>

??? | Diabetes |
??? | Diabetes |
??? | Cancer   |
??? | Diabetes |
??? | Cancer   |
??? | Diabetes |
Example: network access pattern leakage
Example: network access pattern leakage

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td></td>
<td>Diabetes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29489</td>
<td></td>
<td>Diabetes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13744</td>
<td></td>
<td>Cancer</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Learns that Alice has cancer
Goal: oblivious distributed analytics
Goal: **oblivious** distributed analytics

access patterns are independent of data content
Opaque components

Oblivious query planning
- Cost model
- Rule-based opt.
- Cost-based opt.

Distributed oblivious operators
- Oblivious Filter
- Oblivious Aggregation
- Oblivious Join

Computation verification

Data encryption and authentication
Opaque components

Oblivious query planning
- Cost model
- Rule-based opt.
- Cost-based opt.

Distributed oblivious operators
- Oblivious Filter
- Oblivious Aggregation
- Oblivious Join

Computation verification

Data encryption and authentication
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

There can be many partitions
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

Map

Sort

Oblivious sort
[CLRS, Leighton ‘85]
Oblivious aggregation

```sql
SELECT count(*) FROM medical GROUP BY disease
```

<table>
<thead>
<tr>
<th>Count</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Diabetes</td>
</tr>
<tr>
<td>29489</td>
<td>Diabetes</td>
</tr>
<tr>
<td>13744</td>
<td>Cancer</td>
</tr>
<tr>
<td>18740</td>
<td>Diabetes</td>
</tr>
<tr>
<td>98329</td>
<td>Cancer</td>
</tr>
<tr>
<td>32591</td>
<td>Diabetes</td>
</tr>
</tbody>
</table>

Oblivious sort
[CLRS, Leighton '85]
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

Oblivious sort
[CLRS, Leighton ‘85]

Map

Sort

Cancer

Diabetes

13744 ...
98329 ...
12809 ...
29489 ...
18740 ...
32591 ...

Diabetes
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

Map

Sort

Oblivious sort
[CLRS, Leighton '85]
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

Map

Sort

Oblivious sort
[CLRS, Leighton ‘85]
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

<table>
<thead>
<tr>
<th>Diseases</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>98329</td>
</tr>
<tr>
<td>Diabetes</td>
<td>13744</td>
</tr>
<tr>
<td>Diabetes</td>
<td>12809</td>
</tr>
<tr>
<td>Diabetes</td>
<td>29489</td>
</tr>
<tr>
<td>Diabetes</td>
<td>32591</td>
</tr>
<tr>
<td>Diabetes</td>
<td>18740</td>
</tr>
</tbody>
</table>
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

The “Diabetes” group is split!
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

The “Diabetes” group is split!

How to aggregate obliviously and in parallel?
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

Scan
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

Scan

Partial agg.

Boundary processing
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

Scan

Boundary processing
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

Scan

Boundary processing
# Oblivious aggregation

**SQL Query:**

```sql
SELECT count(*) FROM medical GROUP BY disease
```

## Scan

<table>
<thead>
<tr>
<th>ID</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>13744</td>
<td>Cancer</td>
</tr>
<tr>
<td>98329</td>
<td>Cancer</td>
</tr>
<tr>
<td>12809</td>
<td>Diabetes</td>
</tr>
<tr>
<td>29489</td>
<td>Diabetes</td>
</tr>
<tr>
<td>18740</td>
<td>Diabetes</td>
</tr>
<tr>
<td>32591</td>
<td>Diabetes</td>
</tr>
</tbody>
</table>

## Boundary processing

**DUMMY**

- Diabetes: 1
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

Scan

Boundary processing
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

Scan

Boundary processing
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease
Oblivious aggregation

```
SELECT count(*) FROM medical GROUP BY disease
```
SELECT count(*) FROM medical GROUP BY disease
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

Scan  Boundary processing  Scan

13744 ... Cancer
98329 ... Cancer
12809 ... Diabetes

129489 ... Diabetes
18740 ... Diabetes
32591 ... Diabetes

13744 ... Cancer
98329 ... Cancer
12809 ... Diabetes

Diabetes: 1

129489 ... Diabetes
18740 ... Diabetes
32591 ... Diabetes

DUMMY
Cancer: 2
DUMMY
DUMMY
Diabetes: 4
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

Cancer: 2

Diabetes: 4
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease
Oblivious aggregation

```
SELECT count(*) FROM medical GROUP BY disease
```

Oblivious sort
[CLRS, Leighton '85]

Sort
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

[CLRS, Leighton ‘85]
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

Oblivious sort
[CLRS, Leighton ‘85]

Sort

Cancer: 2
Diabetes: 4

Final result
Oblivious aggregation

```
SELECT count(*) FROM medical GROUP BY disease
```

Aggregation has two sorts...

Oblivious sort
[CLRS, Leighton ‘85]

Sort

Final result

Cancer: 2
Diabetes: 4
Oblivious aggregation

SELECT count(*) FROM medical GROUP BY disease

Oblivious sort
[CLRS, Leighton ‘85]

Sort

Aggregation has two sorts…

Can we do better?

Final result

Cancer: 2
Diabetes: 4
Opaque components

Oblivious query planning
- Cost model
- Rule-based opt.
- Cost-based opt.

Distributed oblivious operators
- Oblivious Filter
- Oblivious Aggregation
- Oblivious Join

Computation verification

Data encryption and authentication
Opaque components

Oblivious query planning
- Cost model
- Rule-based opt.
- Cost-based opt.

Distributed oblivious operators
- Oblivious Filter
- Oblivious Aggregation
- Oblivious Join

Computation verification

Data encryption and authentication
Rule-based optimization
Rule-based optimization

```
SELECT count(*)
FROM medical
WHERE age > 30
GROUP BY disease
```
Rule-based optimization

```
SELECT count(*)
FROM medical
WHERE age > 30
GROUP BY disease
```
Insight 1
Insight 1

1. Split each logical operator into smaller Opaque operators
Insight 1

1. Split each logical operator into smaller Opaque operators

2. Take a global view across the plan to remove some Opaque operators
Rule-based optimization

Logical op.

- Aggregation
- Filter
- medical
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

medical
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

medical

medical
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

medical

medical
Rule-based optimization

Opaque op.

Logical op.

- Aggregation
  - Filter
    - medical
- Scan
  - medical
# Rule-based optimization

## Opaque op.

### Logical op.

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
</tr>
<tr>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
</tr>
<tr>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
</tr>
<tr>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
</tr>
<tr>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
</tr>
<tr>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
</tr>
</tbody>
</table>
Rule-based optimization

Opaque op.

Logical op.

Filter

Aggregation

medical

Filter

O-sort

Project

Scan

medical

Rule-based optimization

12809 Amanda D. Edwards 40 Diabetes
129489 Robert R. McGowan 56 Diabetes
13744 Kimberly R. Seay 51 Cancer
18740 Dennis G. Bates 32 Diabetes
32591 Donna R. Bridges 26 Diabetes
98329 Ronald S. Ogden 53 Cancer
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

O-sort

Project

Scan

medical

Filter

medical

Amanda D. Edwards  40  Diabetes
12809

Robert R. McGowan  56  Diabetes
12949

Kimberly R. Seay  51  Cancer
13744

Dennis G. Bates  32  Diabetes
18740

Donna R. Bridges  26  Diabetes
32591

Ronald S. Ogden  53  Cancer
98329
Rule-based optimization

Opaque op.

Logical op.

1.2809 Amanda D. Edwards 40 Diabetes
1.29489 Robert R. McGowan 56 Diabetes
1.3744 Kimberly R. Seay 51 Cancer
1.18740 Dennis G. Bates 32 Diabetes
1.32591 Donna R. Bridges 26 Diabetes
1.98329 Ronald S. Ogden 53 Cancer

Aggregation

Filter

O-sort

Project

Scan

medical

medical
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

O-sort

Project

Scan

medical

medical

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
</tr>
</tbody>
</table>
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

medical

Filter

O-sort

Project

Scan

medical

<table>
<thead>
<tr>
<th>Order</th>
<th>Name</th>
<th>Age</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
</tr>
<tr>
<td>12949</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
</tr>
</tbody>
</table>
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

O-sort

Project

Scan

medical

Filter

O-sort

Project

Scan

medical

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Condition</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>12949</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
<td>1</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
</tbody>
</table>
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

O-sort

Project

Scan

medical

Filter

medical

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Condition</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
<td>1</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
</tbody>
</table>
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

O-sort

Project

Scan

medical

Filter

medical

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Disease</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
<td>1</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
</tbody>
</table>
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

O-sort

Project

Scan

medical

Filter

medical

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
</tr>
<tr>
<td>129489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
</tr>
</tbody>
</table>
# Rule-based optimization

## Logical op.

### Aggregation

### Filter

### Project

### Scan

### medical

## Opaque op.

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
</tr>
</tbody>
</table>
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

O-sort

Project

Scan

medical

Filter

medical

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Diagnosis</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
<td>1</td>
</tr>
</tbody>
</table>
Rule-based optimization

Opaque op.

Logical op.

Aggregation

↑

Filter

↑

O-sort

↑

Project

↑

Scan

↑

medical

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Diagnosis</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
<td>1</td>
</tr>
</tbody>
</table>
Rule-based optimization

Opaque op.

Logical op.

- Filter
- O-sort
- Project
- Scan

Opaque op.

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Diagnosis</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
<td>1</td>
</tr>
</tbody>
</table>
Rule-based optimization

Opaque op.

Logical op.

<table>
<thead>
<tr>
<th>Rule-based optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter</td>
</tr>
<tr>
<td>O-sort</td>
</tr>
<tr>
<td>Project</td>
</tr>
<tr>
<td>Scan</td>
</tr>
<tr>
<td>medical</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rule-based optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rule-based optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td>medical</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rule-based optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809 Amanda D. Edwards 40 Diabetes 0</td>
</tr>
<tr>
<td>129489 Robert R. McGowan 56 Diabetes 0</td>
</tr>
<tr>
<td>13744 Kimberly R. Seay 51 Cancer 0</td>
</tr>
<tr>
<td>18740 Dennis G. Bates 32 Diabetes 0</td>
</tr>
<tr>
<td>98329 Ronald S. Ogden 53 Cancer 0</td>
</tr>
<tr>
<td>32591 Donna R. Bridges 26 Diabetes 1</td>
</tr>
</tbody>
</table>
# Rule-based optimization

## Logical op.

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Diagnosis</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
</tbody>
</table>

## Opaque op.

- Aggregation
- Filter
- O-sort
- Project
- Scan

*medical*
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

medical

Filter

O-sort

Project

Scan

medical

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Condition</th>
<th>Visits</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>129489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
</tbody>
</table>
Rule-based optimization

Opaque op.

Logical op.

Aggregation

O-sort

Agg.

O-sort

Filter

O-sort

Project

Scan

medical

medical

12809  Amanda D. Edwards  40  Diabetes  0
129489 Robert R. McGowan  56  Diabetes  0
13744  Kimberly R. Seay  51  Cancer  0
18740  Dennis G. Bates  32  Diabetes  0
98329  Ronald S. Ogden  53  Cancer  0
Rule-based optimization

Opaque op.

Logical op.

O-sort

Agg.

O-sort

Filter

O-sort

Project

Scan

medical

medical

12809 Amanda D. Edwards 40 Diabetes 0
129489 Robert R. McGowan 56 Diabetes 0
13744 Kimberly R. Seay 51 Cancer 0
18740 Dennis G. Bates 32 Diabetes 0
98329 Ronald S. Ogden 53 Cancer 0
# Rule-based optimization

## Logical op.

- **Aggregation**
  - **Filter**
    - **O-sort**
      - **Filter**
        - **O-sort**
          - **Project**
            - **Scan**
              - **medical**

## Opaque op.

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Diagnosis</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>129489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
</tbody>
</table>
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

medical

O-sort

Agg.

O-sort

Filter

Project

Scan

medical

<table>
<thead>
<tr>
<th>Rule-based optimization</th>
<th>medical</th>
<th>medical</th>
</tr>
</thead>
<tbody>
<tr>
<td>13744 Kimberly R. Seay</td>
<td>51</td>
<td>Cancer 0</td>
</tr>
<tr>
<td>98329 Ronald S. Ogden</td>
<td>53</td>
<td>Cancer 0</td>
</tr>
<tr>
<td>129489 Robert R. McGowan</td>
<td>56</td>
<td>Diabetes 0</td>
</tr>
<tr>
<td>18740 Dennis G. Bates</td>
<td>32</td>
<td>Diabetes 0</td>
</tr>
<tr>
<td>12809 Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes 0</td>
</tr>
</tbody>
</table>
Rule-based optimization

Can we remove any sort?
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

medical

O-sort

Agg.

O-sort

Filter

Project

Scan

medical
Rule-based optimization

Opaque op.

Logical op.

Aggregation

↑

Filter

↑

medical

↑

O-sort

Agg.

O-sort

Filter

O-sort

Project

Scan

medical
Rule-based optimization

Opaque op.

Logical op.

Aggregation

↑

Filter

↑

medical

O-sort

Agg.

O-sort

Filter

O-sort

Project

Scan

medical
Rule-based optimization

Opaque op.

Logical op.

Aggregation

↑

Filter

↑

medical

↑

O-sort

↑

Agg.

↑

O-sort

↑

Filter

↑

Project

↑

Scan

↑

medical

→ Sort on 0/1 column
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

medical

O-sort

Agg.

O-sort

Filter

Project

Sort on 0/1 column

Scan

medical
Rule-based optimization

Opaque op.

Logical op.

Aggregation

↑

Filter

↑

medical

medical

Scan

Project

O-sort

Agg.

O-sort

Filter

Sort on Disease

Sort on 0/1 column
Rule-based optimization

Opaque op.

Logical op.

- Aggregation
- Filter
- medical

- O-sort
- Agg.
- O-sort
- Filter
- Project
- Scan
- medical

Sort on Disease + Sort on 0/1 column
Rule-based optimization

Opaque op.

Logical op.

Aggregation → O-sort → Agg. → O-sort → Filter → O-sort → Project → Scan → medical

Sort on Disease + Sort on 0/1 column =
Rule-based optimization

Opaque op.

Logical op.

Sort on Disease
+
Sort on 0/1 column
=  
Sort on (0/1, Disease)
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

medical

O-sort

Agg.

Filter

O-sort

Project

Scan

medical
Rule-based optimization

Opaque op.

Logical op.

- Aggregation
  - Filter
    - medical

- O-sort
  - Agg.
    - Filter
      - O-sort
        - Project
          - Scan
            - medical
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

medical

O-sort

Agg.

Filter

O-sord

Project

Scan

medical
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

medical

O-sort

Agg.

Filter

O-sort

Project

Scan

medical

medical
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

medical

O-sort

Agg.

Filter

O-sort

Project

Scan

medical
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

medical

O-sort

Agg.

Filter

O-sort

Project

Scan

medical

12809 Amanda D. Edwards 40 Diabetes
129489 Robert R. McGowan 56 Diabetes
13744 Kimberly R. Seay 51 Cancer
18740 Dennis G. Bates 32 Diabetes
32591 Donna R. Bridges 26 Diabetes
98329 Ronald S. Ogden 53 Cancer
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

O-sort

Agg.

Filter

O-sort

Project

Scan

medical

medical

12809 Amanda D. Edwards 40 Diabetes
129489 Robert R. McGowan 56 Diabetes
13744 Kimberly R. Seay 51 Cancer
18740 Dennis G. Bates 32 Diabetes
32591 Donna R. Bridges 26 Diabetes
98329 Ronald S. Ogden 53 Cancer
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

O-sort

Agg.

Filter

O-sort

Project

Scan

medical

medical

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
</tr>
</tbody>
</table>
## Rule-based optimization

### Logical op.

- **Aggregation**
- **Filter**

### Opaque op.

- **O-sort**
- **Agg.**
- **Filter**
- **O-sort**
- **Project**
- **Scan**

### Table

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Diagnosis</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
<td>1</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
</tbody>
</table>
Rule-based optimization

Opaque op.

Logical op.

Aggregation

Filter

medical

O-sort

Agg.

Filter

O-sort

Project

Scan

medical

multi-column sort

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Diagnosis</th>
<th>Authority</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>129489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
<td>1</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
</tbody>
</table>
Rule-based optimization

Opaque op.

Logica op.

Filter

Aggregation

O-sort

Agg.

Filter

O-sort

Project

Scan

medical

medical

multi-column sort

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Condition</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
<td>1</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
</tbody>
</table>
Rule-based optimization

Logical op.

Opaque op.

O-sort
Agg.
Filter
O-sort
Project
Scan
medical

medical

multi-column sort

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Age</th>
<th>Condition</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
<td>1</td>
</tr>
</tbody>
</table>
Rule-based optimization

Logical op.

Opaque op.

O-sort
Agg.
Filter
O-sort
Project
Scan

multi-column sort

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Disease</th>
<th>l0</th>
</tr>
</thead>
<tbody>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
<td>1</td>
</tr>
</tbody>
</table>
Rule-based optimization

Logical op.

Opaque op.

multi-column sort

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Disease</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>32591</td>
<td>Donna R. Bridges</td>
<td>26</td>
<td>Diabetes</td>
<td>1</td>
</tr>
</tbody>
</table>

Opaque op.

Filter

Agg.

O-sort

Project

Scan

medical

Filter

Aggregation

medical
Rule-based optimization

Logical op.

Opaque op.

O-sort

Agg.

Filter

O-sort

Project

Scan

medical

multi-column sort

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Diagnosis</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
</tbody>
</table>
Rule-based optimization

Logical op.

Opaque op.

O-sort

Agg.

Filter

O-sort

Project

Scan

medical

medical

multi-column sort

Eliminated one oblivious sort!

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
<th>Disease</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>13744</td>
<td>Kimberly R. Seay</td>
<td>51</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>98329</td>
<td>Ronald S. Ogden</td>
<td>53</td>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>12809</td>
<td>Amanda D. Edwards</td>
<td>40</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>18740</td>
<td>Dennis G. Bates</td>
<td>32</td>
<td>Diabetes</td>
<td>0</td>
</tr>
<tr>
<td>29489</td>
<td>Robert R. McGowan</td>
<td>56</td>
<td>Diabetes</td>
<td>0</td>
</tr>
</tbody>
</table>
Rule-based optimization

Eliminated one oblivious sort!

No change to the Spark SQL planner!
More in the paper..

Oblivious query planning
- Cost model
- Rule-based opt.
- Cost-based opt.

Distributed oblivious operators
- Oblivious Filter
- Oblivious Aggregation
- Oblivious Join

Computation verification

Data encryption and authentication
Evaluation setup
Evaluation setup

- Single machine experiments:
  - Intel Xeon E3-1280 v5, 4 cores, 64 GB RAM
  - Intel SGX: 128 MB of enclave page cache (EPC)
Evaluation setup

- Single machine experiments:
  - Intel Xeon E3-1280 v5, 4 cores, 64 GB RAM
  - Intel SGX: 128 MB of enclave page cache (EPC)
- Distributed experiments
  - A cluster of 5 SGX machines
Evaluation
Evaluation

- How does Opaque compare to Spark SQL?
Evaluation

• How does Opaque compare to Spark SQL?
  • Big Data Benchmark (BDB); 4 queries total
Evaluation

• How does Opaque compare to Spark SQL?
  • Big Data Benchmark (BDB); 4 queries total
    • Queries 1, 2, 3: filter, aggregation, join
Evaluation

• How does Opaque compare to Spark SQL?
  • Big Data Benchmark (BDB); 4 queries total
    • Queries 1, 2, 3: filter, aggregation, join
    • 1 million records
Evaluation

• How does Opaque compare to Spark SQL?
  • Big Data Benchmark (BDB); 4 queries total
    • Queries 1, 2, 3: filter, aggregation, join
    • 1 million records

• How does Opaque compare to state-of-the-art oblivious systems?
Evaluation

• How does Opaque compare to Spark SQL?
  • Big Data Benchmark (BDB); 4 queries total
    • Queries 1, 2, 3: filter, aggregation, join
    • 1 million records
• How does Opaque compare to state-of-the-art oblivious systems?
  • GraphSC (oblivious graph analytics)
Evaluation

• How does Opaque compare to Spark SQL?
  • Big Data Benchmark (BDB); 4 queries total
    • Queries 1, 2, 3: filter, aggregation, join
    • 1 million records

• How does Opaque compare to state-of-the-art oblivious systems?
  • GraphSC (oblivious graph analytics)
    • PageRank
Big Data Benchmark (distributed)
Big Data Benchmark (distributed)

Data encryption, authentication, computation verification
## Big Data Benchmark (distributed)

Data encryption, authentication, computation verification

<table>
<thead>
<tr>
<th>Query number</th>
<th>Spark SQL</th>
<th>Opaque</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query 1</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Query 2</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Query 3</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Runtime (s)

- 100
- 10
- 1
- 0.1
- 0.01
Big Data Benchmark (distributed)

Data encryption, authentication, computation verification

- Spark SQL
- Opaque

Runtime (s)

Query number

- Query 1
- Query 2
- Query 3
Big Data Benchmark
(distributed)

Data encryption, authentication, computation verification

Runtime (s)

Query 1  Query 2  Query 3

Query number

Spark SQL  Opaque
Big Data Benchmark (distributed)

Data encryption, authentication, computation verification

Overhead: -0.47x to 2.3x
Big Data Benchmark (distributed)

Data encryption, authentication, computation verification

+ Obliviousness

Overhead: -0.47x to 2.3x
Big Data Benchmark (distributed)

Data encryption, authentication, computation verification

Runtime (s)

Query number

Spark SQL
Opaque

+ Obliviousness

Query number

Runtime (s)

Query 1  Query 2  Query 3

Overhead: -0.47x to 2.3x
Big Data Benchmark (distributed)

Data encryption, authentication, computation verification

+ Obliviousness

Overhead: -0.47x to 2.3x
Big Data Benchmark (distributed)

Data encryption, authentication, computation verification

+ Obliviousness

Query number

Overhead: -0.47x to 2.3x
Big Data Benchmark (distributed)

Data encryption, authentication, computation verification

Run-time (s)

Query number

Spark SQL
Opaque

Overhead: -0.47x to 2.3x

+ Obliviousness

Run-time (s)

Query number

Spark SQL
Opaque

Overhead: 21x to 45x
PageRank: comparison with GraphSC (single machine)
Conclusion

Opaque is an oblivious and encrypted distributed analytics platform

WIP open source: github.com/ucbrise/opaque