Homework 5 | Query Execution & Assessing Societal Impacts

Updates made to the assignment spec after release are highlighted in red.

Objectives: To translate fluidly from SQL to RA. To understand the statistics used by SQL engines and how they inform modifications to RA trees. To identify the likelihood of harming society using the impact/scope/opacity rubric.

Due date: Thu, May 15 2025 @ 9:00pm

***Median completion time (24sp):* 4 hours**

Resources

For this assignment, you will need:

- Pen and paper, or any drawing tools you prefer (e.g., PowerPoint, draw.io).

Problem Set

The technical questions (i.e., all questions except the last three) refer to the following tables:

```
CREATE TABLE Customers (
    id          INTEGER PRIMARY KEY,
    firstname   VARCHAR(64) NOT NULL,
    lastname    VARCHAR(64) NOT NULL,
    city        VARCHAR(64) NOT NULL
);
CREATE TABLE Orders (
    id          INTEGER PRIMARY KEY,
    cid         INTEGER REFERENCES Customers,
    orderDate   DATE      -- assume this type is equivalent to
                          -- INTEGER; ie >, =, MAX(), etc are
                          -- well-defined
```

```
);
CREATE TABLE OrderItems (
    oid          INTEGER REFERENCES Orders,
    itemName     VARCHAR(64) NOT NULL,
    quantity     INTEGER,  -- guaranteed >0
    isFulfilled  INTEGER   -- boolean
);
```

1. *(this blank question ensures that this document's numbering matches Gradescope's numbering of questions)*

2. The following queries are semantically identical to the queries from section, though some names have changed. Note that they use bind variables, which we have not covered in class; however, from an index-design perspective you can assume they behave as if they were a constant value (eg, "2024-05-09").

   This query is executed several times per second:

```
-- Identify the items we need to put into a shipment (any shipment)
SELECT oi.itemName, SUM(oi.quantity)
FROM OrderItems oi
WHERE oi.isFulfilled = 0
GROUP BY oi.itemName;
```

   This query is executed once per hour:

```
-- Identify the lingering items ordered by the Bezos family which have
-- not been shipped yet, and the city to which they'd like their items to
-- be shipped.
SELECT oi.itemName, c.city, SUM(oi.quantity)
FROM Orders o, OrderItems oi, Customers c
WHERE o.id = oi.oid AND o.cid = c.id
AND oi.isFulfilled = 0
AND c.lastName = 'Bezos'
AND o.orderDate < ?
GROUP BY oi.itemName, c.city
ORDER BY o.orderDate, SUM(oi.quantity);
```

   These queries are executed approximately once every second:

```
-- Record a new order
INSERT INTO Orders VALUES (?, ?, 0, ?);
```

```
-- Record their ordered items.  An "average" order has 3-5 items, but
-- it is possible to have a single-item order.  Orders are capped to 256
-- unique items, enforced by the database application.
INSERT INTO OrderItems VALUES (?, ?, ?);
INSERT INTO OrderItems VALUES (?, ?, ?);
INSERT INTO OrderItems VALUES (?, ?, 0, ?);
-- ... etc ...
```

This query is executed many many times per second, but only during business hours:

```
-- Ship an item
UPDATE OrderItems
SET isFulfilled = 1
WHERE oid = ?
AND itemName = ?;
```

Lastly, approximately 10% of queries are none of these; they're a mixture of tools that aren't run frequently or adhoc reports run by analysts looking for interesting patterns (maybe brand loyalty to Bose earbuds?). We don't want to optimize our database for these queries.

What indices would you create to support this query load? Unlike the section worksheet, we would like you to *consider the query load holistically* rather than considering each query individually. For example, if you decide to have more than one index on a table you must also describe how you chose which index to cluster.

You can assume the DBMS has **not** created any indices, not even an index on the tables' primary keys.

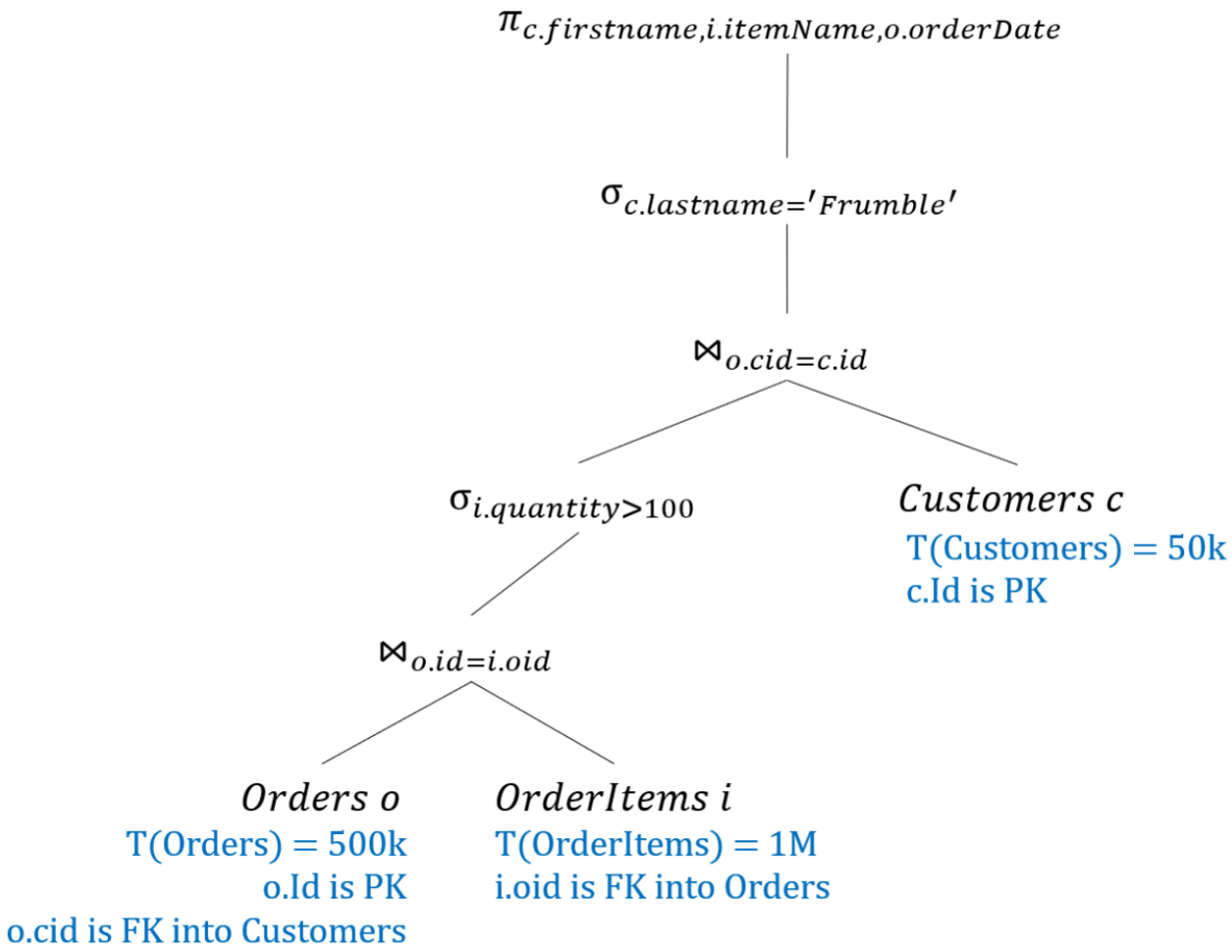
3. Convert the following SQL query to a relational algebra tree. If relevant, please:
    - Draw the tables from left to right in the same order in which they appear in the FROM clause
    - Order any joins in the same order in which they appear in the query

```
  SELECT o.cid
    FROM Orders o, OrderItems i
   WHERE o.id = i.oid
     AND (i.itemName = 'Bose QuietComfort Earbuds'
          OR i.itemName = 'Beats Fit Pro')
GROUP BY o.cid
  HAVING COUNT(DISTINCT i.itemName) = 2;
```

4. This and the next two problems require the following statistics:

T(Customers) = 50,000 T(Orders) = 500,000 T(OrderItems) = 1,000,000 V(Customers, lastname) = 5000 min(OrderItems, quantity) = 1 max(OrderItems, quantity) = 200

and refer to this RA tree:



For each of the 5 stages, determine its selectivity factor. Then, compute the cardinality of those stages; as a hint, the final projection produces 100 tuples.

5. Starting from the same tree as question 4, push the selection operators down as low as possible. Submit your modified tree to Gradescope.

6. For each of the 5 stages *in the tree you modified submitted for question 5*, determine its selectivity factor and cardinality. Then, compute the cardinality of those stages; the final projection will still produce 100 tuples.

7. Consider the trees (and their cardinality estimates) from question 4 and question Which tree executes more efficiently? Why?

8. Please read the following two sections:
    - How we rank Feed and Stories
    - How you can influence what you see

    of this article describing Facebook's algorithm for ranking Instagram posts. Then, describe the scale, impact, and opacity of this algorithm. You should be able to answer each section in a maximum of 3-5 sentences, and you may find this example helpful.

9. Next, please reflect on your own personal experience. Specifically:
    - What is one thing that you *learned* while doing this assignment?
    - What is one thing that *surprised you* while doing this assignment?
    - What is one *question that you still have* after doing this assignment?

10. Lastly, please answer the feedback questions:
    - How many hours did it take you to finish this assignment, including time to set up your computer (if necessary)?
    - How many of those hours did you feel were valuable and/or contributed to your learning?
    - Did you collaborate with other students on this assignment? If so, approximately how many people did you collaborate with? Do not include yourself or course staff in the count.

Submission Instructions

You should submit the questions on Gradescope under HW5. You will need to submit two different RA trees, which we recommend uploading in PDF format.