Database Systems
CSE 414

Section 10: Big Data & Review

Non-Parallel Query Evaluation

Example Schema

Product(pid, name, category)

- 10,000 tuples and 1,000 blocks
- 40 different categories

Order(store, pid, price, quantity)

- 1,000,000 tuples and 50,000 blocks
- prices range from $1 to $100

Example Query

Compute the total revenue, for each store, from electronics costing more than $5 each:

```
SELECT o.store, sum(o.price * o.quantity)
FROM Order o, Product p
WHERE o.pid = p.pid AND o.price > 5 AND
   p.category = 'electronics'
GROUP BY o.store
```

Problem 1

Give an RA expression that:

- computes the result of the query
- **does not** benefit from the indexes already present

Problem 2

Estimate the cost of the RA expression from Problem 1 after filling in physical implementation details

- assume grouping / aggregation can be done on the fly

Problem 3

Give an RA expression that:

- computes the result of the query
- **does** benefit from the indexes already present

Problem 4

Estimate the cost of the RA expression from Problem 3 after filling in physical implementation details

- assume grouping / aggregation can be done on the fly

Parallel Query Evaluation

Problem 5

Draw a pipeline that computes the same result in a parallel fashion using N nodes

Problem 6

Estimate the cost of executing the pipeline of Problem 5

Problem 7

1. Does your analysis predict a linear speedup as more nodes are added?

2. Does your analysis predict a linear scaleup as more nodes are added?

3. How realistic is this?

Problem 8

Describe how to achieve a similar speedup with MapReduce

Problem 9

Would your MapReduce have the same IO cost and speedup as the pipeline from problem 6?