

Introduction to Database Design, Fall 2012

Rasmus Pagh

November 29, 2012

To give an impression of what may come, the below problem is of a form that could be given at the written exam.

MapReduce data processing (10%)

Suppose you have a traditional DBMS that contains a table

```
Purchase(id,date,shopId,sellerId,category,itemId,salesPrice)
```

and every day runs the following query:

```
SELECT shopId, sum(salesPrice)
FROM Purchase
WHERE category = 'food'
GROUP BY shopId
```

Since the amount of data is getting large you consider placing the data in a file (one tuple per row), and process it using MapReduce.

a) Describe map and reduce functions that would implement the above query. You may use pseudo-code, or give a description in words.

b) Describe map and reduce functions that will be able to compute the most sold itemId in each category (in terms of number of sales). More than one MapReduce computation may be needed. You may use pseudo-code, or give a description in words.