CSE-444: Homework #4
Due: Wednesday, March 11th

1. Ex. 6.4.3, (a,b) – Page 347.

2. Ex. 6.6.2, (a,b,c) – Page 357.

3. Consider a relation $R(a, b, c, d)$ containing 5,000,000 tuples, where each data page of the relation holds 10 records. R is stored as a sorted file on the attribute a, which is also a key of R. R also has secondary indexes. The values of $R.a$ range from 0 to $4,999,999$. For each of the following relational algebra queries, state which of the following three approaches is most likely to be the cheapest:

- Access the sorted file for R directly,
- Use a clustered B+ inex on the attribute $R.a$,
- Use a linear hashed index on attribute $R.a$.

1. $\sigma_{a<50,000}(R)$
2. $\sigma_{a=50,000}(R)$
3. $\sigma_{a>50,000 \wedge a<50,010}(R)$
4. $\sigma_{a\neq 50,000}(R)$

Remember to state any additional assumptions you may be making.

4. Consider the following schema:

$Emp(\underline{eid : \texttt{integer}}, sal : \texttt{integer}, age : \texttt{real}, did : \texttt{integer})$
$Dept(\underline{did : \texttt{integer}, projid : \texttt{integer}}, budget : \texttt{real}, status : \texttt{char(10)})$
$Proj(\underline{projid : \texttt{integer}}, code : \texttt{integer}, report : \texttt{varchar})$

Assume that each *Emp* record is 20 bytes long, each *Dept* record is 40 bytes long and each *Proj* record is 2000 bytes long on average. There are 20,000 tuples in *Emp*, 5000 tuples in *Dept* and 1000 tuples in *Proj*. Note that *did* is *not* a key of *Dept*. We assume our page size is 4000 bytes, and that there are 12 pages available in the buffer pool. You can assume uniform distribution of values whenever necessary.

Consider the query:

```
SELECT *
FROM Emp E, Dept D
WHERE E.did = D.did
```

Suppose that there is a clustered hash index on *did* of *Emp*. List all the plans that are considered and identify the plan with the least cost.

Note:

1. You need to consider the following join algorithms: block nested-loop join, hash join, and sort-merge join.
2. The cost metric of plans is the number of I/O operations. Ignore the cost of writing out the final result.
3. Remember to make explicit any other assumptions you are making.