Quiz 3: Query Processing and Query Optimization

A: THE YELP DATABASE [70pts]

Suppose we are given a database with the following schema.

**Users** (UserID INTEGER, Name CHAR(30), Age INTEGER, ReviewCount INTEGER)

**Businesses** (BusinessID INTEGER, BName CHAR(30), City CHAR(20), State CHAR(2))

**Checkins** (BusinessID INTEGER, Weekdays INTEGER, Weekends INTEGER)

**Reviews** (ReviewID INTEGER, UserID INTEGER, BusinessID INTEGER, Stars REAL)

**Reviews** (UserID) is a foreign key referring to **Users** (UserID).
**Reviews** (BusinessID) is a foreign key referring to **Businesses** (BusinessID).
**Checkins** (BusinessID) is a foreign key referring to **Businesses** (BusinessID).

A page is 8 kB in size. The RDBMS buffer pool has 10,000 pages, all of which are usable. Initially, the buffer pool is empty.

The relation instances have the following statistics. Assume there are no NULL values. Each integer or real is 8B, and each character is 1B (so as an example CHAR(20) is 20B). Additionally, the record id of each tuple is 8B.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Number of Pages</th>
<th>Number of Tuples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>75,684</td>
<td>10m</td>
</tr>
<tr>
<td>Businesses</td>
<td>41,504</td>
<td>5m</td>
</tr>
<tr>
<td>Checkins</td>
<td>19,532</td>
<td>5m</td>
</tr>
<tr>
<td>Reviews</td>
<td>488,282</td>
<td>100m</td>
</tr>
</tbody>
</table>

Answer the following questions. *Clearly explain how you obtained your answer for each.*

1. **[10pts]** Name 5 indexes (hash and/or clustered B+ tree) on Users that match the predicate in the following SQL query and explain why each index matches.

```sql
SELECT *
FROM Users
WHERE NOT ((Name <> "John" AND NOT (Name = "Mary"))
OR (Age <> 20 AND Age <= 50));
```
2. [10pts] Suppose we are given a clustered B+ tree index on Businesses (State, City) with a fan-out of 100. Also, suppose that the index follows the alternative of storing the data records directly in the leaf pages of the index. What is the I/O cost (number of page I/Os) of the following SQL query? Exclude the cost of writing the output. Assume that the selectivity of the predicate State = “WI” is 2%. Show all of your calculations clearly.

```
SELECT DISTINCT City
FROM Businesses
WHERE State = "WI";
```

3. [15pts] Suppose we are given a clustered B+ tree index each on Businesses (BusinessID) and Checkins (BusinessID), both with a fan-out of 100. Also, suppose that both indexes follow the alternative of storing the data records directly in the leaf pages of the index. Which join algorithm among the following has the lowest I/O cost for a natural join of Businesses and Checkins: Block Nested Loop Join, Index Nested Loop Join, Sort-Merge Join, or Hash Join? Show all of your calculations clearly.

4. [10pts] Suppose that there is no index on the Businesses relation. Consider the following SQL query.

```
SELECT City, COUNT (BusinessID)
FROM Businesses
GROUP BY City;
```

What is the maximum number of cities for which it is possible to implement hash-based aggregation by reading the relation only once? Assume that the fudge factor of the hash table is \( f = 1.4 \). Show all of your calculations clearly.

5. [10pts] Suppose that there are no indexes on any relation and no relation is sorted on any attribute. Propose a physical plan for the following SQL query that does not materialize any intermediate relation, and compute its I/O cost. Assume that the values of Stars are real numbers uniformly distributed between 0 and 5 (inclusive), and the values of Age are integers uniformly distributed between 10 and 99 (inclusive). Also assume independence of the predicates on Stars and Age. Show all of your calculations clearly.

```
SELECT COUNT (UserID)
FROM Users U, Reviews R
WHERE U.UserID = R.UserID AND R.Stars < 1 AND U.Age = 18;
```
6. **[15pts]** Suppose we are given a hash index on Users (Age), and suppose that the index follows the alternative of storing the data records on a different clustered file. Propose a physical plan for the following SQL query and compute its I/O cost. Your plan must use the available hash index, and include at every operator whether it uses materialization or pipelining. The same assumptions as the previous question hold for the distribution of the values of Stars and Age. Also assume the independence of the two predicates. Show all of your calculations clearly.

```sql
SELECT COUNT (DISTINCT UserID)
FROM Users U, Reviews R
WHERE U.UserID = R.UserID AND R.Stars < 1 AND U.Age = 18;
```

**B: MISCELLANEOUS [30pts]**

1. **[15pts]** Hash Join with non-uniform partitioning:

   Suppose we are joining two tables S and R with respective number of pages $4BN_S$ and $8BN_R$, wherein $4BN_S \gg 8BN_R$. The number of buffer pages is $4B + 1$ and the buffer pool is initially empty. We are also given that $2fN_R = 4B - 1$, where $f$ is the hash table fudge factor.

   The distribution of the join attribute values in S and R are such that after the first hash partitioning phase, we get exactly $4B$ partitions of S, each of length $N_S$ pages, but not all partitions of R are of the same length. Suppose R gets partitioned as follows: $2B$ partitions of length $N_R$ pages, $B$ partitions of length $2N_R$ pages, and $B$ partitions of length $4N_R$ pages.

   What is the I/O cost of the regular hash join algorithm discussed in class? Exclude the cost of writing the output of the join. Assume perfect uniform splitting occurs during the recursive repartitioning. Show all of your calculations clearly.

   (Hint: The answer is of the following form: $xBN_S + yBN_R$, where $x \in \{12, 14, 16, 18\}$ and $y \in \{24, 28, 32, 36\}$.)

2. **[15pts]** Query Rewriting:

   Consider the following query expressed in Relational Algebra over the Yelp database given in Part A:

   $$\pi_{BName}(\sigma_{Stars>4,ReviewCount\geq100}((Users \bowtie Reviews) \bowtie Businesses))$$

   Write an equivalent Relational Algebra expression where the selections and projections are pushed down the plan as far as possible. Apply only the rules from Chapter 15.3.4 of the Cow Book. Show and explain each step in your rewriting.
C: EXTRA CREDIT (OPTIONAL) [+20pts]

1. [10pts] Consider the Yelp database given in Part A. What is the I/O cost of bulk loading a clustered B+ tree index on Reviews (Stars)? Assume that we follow the alternative of storing the data records directly in the leaf pages of the index. Also, assume that all non-leaf pages are kept in buffer memory during the bulk loading process, but are all written out to disk once at the end of the process. Include this cost of persisting the index. Show all of your calculations clearly.

2. [10pts] Let $R, S, T$ be three relations such that $S$ and $T$ have the same schema. Consider the following equation:

$$R \bowtie_{\theta} (S - T) = (R \bowtie_{\theta} S) - (R \bowtie_{\theta} T)$$

where $\theta$ is a join condition. Is the above identity true? If yes, provide a proof. If no, give a counterexample that falsifies the identity.