SJSU Students
CS157B-01
HW 3

Question 1
For each of the following operations, write an iterator that uses an algorithm described in class to enumerate the output of the following operations: (a) project tuples onto attributes A,B,C, renaming C to D, (b) bag union. Make sure your iterator reads whole blocks at a time, but outputs tuples at time.

a) Project tuples onto attributes A, B, C renaming to C to D
   Note that projection is a one-pass, tuple-at-a-time operation.
   1. First, open R. Use GetNext in a loop to the blocks of R.
   2. Use M-1 blocks to store the tuples seen so far.
   3. For each tuple t1 from R, form a new tuple t2 with attributes A, B, and D using values from t1.A, t1.B, and t1.C respectively.
   4. Move t2 into output block (consider this as output buffer)
   5. Output each tuple in the output block
   6. Close R

b) Bag union
   1. First, open R. Use GetNext in a loop to read blocks in R til R is done.
   2. Open S. Use GetNext in a loop to read blocks in S till S is done too.
   3. For each tuple t in S, count the number of occurrences in R and S, and call these counts r and s
   4. Make r + s copies of tuple t and write these copies to the output block
   5. Output the tuples in the output block
   6. Close R and S

Question 2
If B(R)=B(S)=150,000 and M=6000, what are the disk I/O requirements of: (a) two-pass set intersection using hashing, (b) sort-join from class.

M = number of main memory blocks available
B(R) = number of blocks in R
T(R) = number of tuples in R

a) Two-pass set intersection using hashing
   Is B(R) + B(S) < M(M-1)?

150,000 + 150,000 < 6000(6000-1)?
300,000 < 35994000 ? Yes, so

Will take time 3(B(R) + B(S))
3 * (150,000 + 150,000) = **900,000**

b) Sort-join
Is B(R) + B(S) < M^2?
150,000 + 150,000 < (6000)^2 ?
300,000 < 36,000,000 ? Yes, so

Will take time 3(B(R) + B(S))
3 * (150,000 + 150,000) = **900,000**

Question 3

Come up with additional query parsing rules to add to our rules from the Mar 8 Lecture to say what <Relations> and <Tuple> are. Make sure your rules can handle expressions for relations like Employee E.

<Query> ::= <SFW>
<Query> ::= (<Query>)
<SFW> ::= SELECT <SelList> FROM <FromList> WHERE <Condition>
<SelList> ::= <Attribute>, <SelList>
<SelList> ::= <Attribute>
<FromList> ::= <Relation>, <FromList>
<FromList> ::= <Relation>
<Condition> ::= <Condition> AND <Condition>
<Condition> ::= <Tuple> IN <Query>
<Condition> ::= <Attribute> = <Attribute>
<Condition> ::= <Attribute> LIKE <Pattern>

<Relation> ::= <RelName>
<Relation> ::= <RelName> <Alias>
<RelName> ::= [A-Za-z_]+[0-9]
<Alias> ::= [A-Za-z_]+[0-9]

<Tuple> ::= (<AttrList>)
<AttrList> ::= <Attribute>, <AttrList>
<AttrList> ::= <Attribute>

Question 4
Consider the table:

Y(c,d)	Z(d,e)
T(Y)=1200	T(Z)=1400
V(Y,c)=30	V(Z,d)=90
V(Y,d)=70	V(Z,e)=60

Estimate the sizes of relations that are the results from the following queries:
(a) $\sigma_{e=95}(Z)$, (b) $Y \bowtie \sigma_{e=95}(Z)$.

a) $\sigma_{e=95}(Z)$
Size of a Selection: $S = \sigma_{A=c}(R)$ means $T(S) = T(R) / V(R,A)$

In this case, e is our attribute and our relation is Z.
$T(S) = T(Z) / V(Z,e)$
$T(S) = 1400 / 60 =$ **23.333, so about 24 if we round up**

b) $Y \bowtie \sigma_{e=95}(Z)$
Size of a Join: $R(X,Y) \bowtie S(Y,Z)$ means $T(R \bowtie S) = (T(R)*T(S)) / \max(V(R,Y), V(S,Y))$

In this case, Y and Z will be the relation R and S respectively.
Note that $T(\sigma_{e=95}(Z))$ has already been computed in the previous question.

$T(Y \bowtie \sigma_{e=95}(Z)) = (T(Y) * T(\sigma_{e=95}(Z))) / \max(V(Y,d), V(Z,d))$
$T(Y \bowtie \sigma_{e=95}(Z)) = (1200 * 24) / \max(70, 90)$
$T(Y \bowtie \sigma_{e=95}(Z)) = 28800 / 90 =$ **320**

// Question 5 on next page

Question 5

Assume A=20,B=30 (here we imagine A and B are blocks that can hold 1 integer) are stored in a DB. Suppose a transaction does the following sequence of operations I(A), I(B), R(B,b), b:= b+5, R(A,a), W(B,b), a:= 2*a + b, W(A,a), O(A), O(B). Show the undo log records needed for this transaction.

Transaction op	**Value a**	**Value b**	**Mem Value A**	**Mem Value B**	**Disk Value A**	**Disk Value B**	**Log Records**
							<START T>
I(A)			20		20	30	
I(B)			20	30	20	30	
R(B,b)		30	20	30	20	30	
b:=b+5		35	20	30	20	30	
R(A,a)	20	35	20	30	20	30	
W(B,b)	20	35	20	35	20	30	<T, B, 30>
a:=2*a+b	75	35	20	35	20	30	
W(A,a)	75	35	75	35	20	30	<T, A, 20>
FLUSH LOG	75	35	75	35	20	30	
O(A)	75	35	75	35	75	30	
O(B)	75	35	75	35	75	35	
							<COMMIT T>
FLUSH LOG							