Name: _____

This exam is open book, open notes, but there can be no sharing of any material. You can use the Internet, but only as a library. If you are not sure if something is allowed, check with me. Some questions are marked with a ⋆. The points for the non-⋆ questions sum to 80, while those for the ⋆ questions sum to 40.

COS 480 students must answer all questions that are not marked with a ⋆. The points for each question are indicated in parentheses next to the question number. Questions marked with a ⋆ may also be answered, for extra credit.

COS 580 students must answer all questions, including those marked with a ⋆, in 100 minutes. Each question is worth 2/3 times the points indicated in parentheses.

Several questions on this exam use the database instance suggested below. A row in the `PTides` table represents a predicted tide and lists the location, time, kind (high or low), and height of the prediction. A row in the `DockSched` table represents a scheduled arrival of a boat at a dock and lists the harbor, boat's name, pilot's name, scheduled arrival time, and boat's length. Your answers to questions that ask for queries should work for all instances of databases conforming to the given schema, not only the one depicted below.

These tables are repeated on the last page of the test. (You may detach that page and use it for reference. There is no need to reattach it.)

PTides

location varchar(20)	ptime timestamp	kind char(1)	height float
Blue Hill	2005-10-13 01:27	L	0.29
Blue Hill	2005-10-13 07:42	H	9.80
Blue Hill	2005-10-13 13:47	L	1.00
Blue Hill	2005-10-13 19:59	H	10.98
Eastport	2005-10-13 01:25	L	0.77
Eastport	2005-10-13 07:31	H	17.60

DockSched

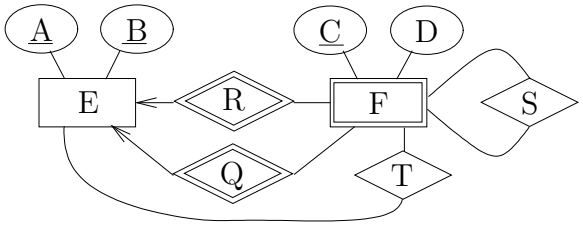
harbor varchar(20)	boat varchar(20)	pilot varchar(20)	dtime timestamp	blength integer
Blue Hill	Why Knot	Knotting	2005-10-13 08:00	14
Blue Hill	Why Knot	Knotting	2005-10-14 08:00	14
Blue Hill	Phair Game	Phair	2005-10-13 08:10	10
Castine	Phair Game	Phair	2005-10-13 08:30	10

For notational convenience in relational algebra, we use the following abbreviations:

`PTides(location, ptime, kind, height)`	$P(L, P, K, H)$
`DockSched(harbor, boat, pilot, dtime, blength)`	$D(H, B, P, D, L)$

1. (1 pt.) Write your name in the space provided above.

2. (4 pts.) What is the primary key of the entity set F in the following E-R diagram?



3. (5 pts) Generate a relational schema that best corresponds to the E-R diagram of Question 2. Underline key attributes.

4. (5 pts.) Consider the relation $R(A, B, C, D, E)$ with the following basis of dependencies:

$$AB \rightarrow CD \quad (1)$$
$$C \rightarrow D \quad (2)$$
$$D \rightarrow AE \quad (3)$$
$$CE \rightarrow AB \quad (4)$$

Compute $\{A, B\}^+$ and $\{B, C, E\}^+$.

5. (10 pts.) Determine all keys of the relation R of Question 4. Justify your answer.

6. (15 pts.) Render the schema of Question 4 in BCNF. *For each decomposition step, you must clearly list (1) the dependency used for the decomposition and (2) a basis of the projected dependencies for the resulting relations.* Summarize the final schema (list all its relations) and explain why it is in BCNF.

7. (5 pts.) Write a SQL statement to express the following constraint: The docking times for each pilot must be at least 30 minutes apart. Use the simplest kind of SQL constraint that is capable of expressing the given constraint.

8. (5 pts.) We say a pilot p *hops* from a harbor s (the source) to a harbor d (the destination) if p is scheduled to dock at a and b at times t_1 and t_2, respectively, with $t_1 < t_2$, and p is not scheduled to dock at any time between t_1 and t_2. Write a SQL query that returns tuples (p, s, d) where pilot p hops from s to d.

9. (5 pts.) Write a relational algebra query that is equivalent to the query of Question 8.

10. (5 pts.) Write a Datalog query that is equivalent to the query of Question 8.

11. (10 pts.) We say a hop (Question 8) is a *high hop* (respectively, *low hop*) if the docking times at both source and destination are within one hour of high (respectively, low) tides. We say a harbor B is *high-hop reachable* (respectively, *low-hop reachable*) from a harbor A if there is a sequence of high (respectively, low) hops $h_1, h_2, \ldots, h_n$, with $n \geq 2$, such that $h_1 = A$, $h_n = B$, and the source of hop h_i equals the destination of hop h_{i-1} for all $i = 2, \ldots, n$. Write a Datalog query for pairs of harbors (x, y) such that y is high-hop reachable from x but y is not low-hop reachable from x.

12. (10 pts.) The following questions are based on Codd's paper in the reading list.

    (a) Provide a relational-algebra expression (using the textbook's version of the algebra) for the *restriction* operation (Section 2.1.5, page 385).

    (b) Write the simplest standard-SQL statement that expresses the constraint described in the second paragraph of Section 2.3 (page 386, right column).

13. (10 pts.) ⋆ We say a harbor B is *alt-hop reachable* from a harbor A if there is a sequence of alternating high and low hops from A to B. (Recall the definitions of Question 11.) More precisely, B is alt-hop reachable from A if there is a sequence of $n \geq 2$ hops $A = h_1, h_2, \ldots, h_n = B$ such that, for all $i = 2, \ldots, n$, (1) the source of hop h_i equals the destination of hop h_{i-1} and (2) either h_i is a high hop and h_{i-1} is a low hop or vice versa. Write a Datalog query for pairs of harbors (x, y) such that y is alt-hop reachable from x.

14. (5 pts.) Using only the six basic operators, provide a relational algebra expression that is equivalent to the following SQL query:

```
select *
from PTides full outer join DockSched on location = harbor
where ptime = dtime;
```

15. (25 pts.) ⋆ The following questions are based on Graefe's paper in the reading list.

    (a) (5 pts.) Table 1 on page 81 indicates that the open function for a sort iterator performs all merge steps except the last one. Comment on the advantages and drawbacks of this design over one that completes all merge steps in the open function.

(b) (10 pts.) Provide a derivation of the formula for determining the fan-in of the first merge (page 88). Do not simply repeat the outline in the paper; fill in all the missing steps.

(c) (10 pts.) Provide pseudocode for the block nested-loop join operator. Assuming a LRU buffer-replacement policy, provide an expression for the number of I/O operations when joining relations of sizes N_1 and N_2 bytes with M bytes of RAM and a cluster size of C. Assume that M is substantially smaller than N_1 and N_2.

Scratch page

Material here will not be graded. You may detach and discard this page.

PTides

location varchar(20)	ptime timestamp	kind char(1)	height float
Blue Hill	2005-10-13 01:27	L	0.29
Blue Hill	2005-10-13 07:42	H	9.80
Blue Hill	2005-10-13 13:47	L	1.00
Blue Hill	2005-10-13 19:59	H	10.98
Eastport	2005-10-13 01:25	L	0.77
Eastport	2005-10-13 07:31	H	17.60

DockSched

harbor varchar(20)	boat varchar(20)	pilot varchar(20)	dtime timestamp	blength integer
Blue Hill	Why Knot	Knotting	2005-10-13 08:00	14
Blue Hill	Why Knot	Knotting	2005-10-14 08:00	14
Blue Hill	Phair Game	Phair	2005-10-13 08:10	10
Castine	Phair Game	Phair	2005-10-13 08:30	10

For notational convenience in relational algebra, we use the following abbreviations:

`PTides(location, ptime, kind, height)`	$P(L, P, K, H)$
`DockSched(harbor, boat, pilot, dtime, blength)`	$D(H, B, P, D, L)$