CS150A–Database
Final Exam Solutions

January 7, 2022

I Basics and SQL [10 points]

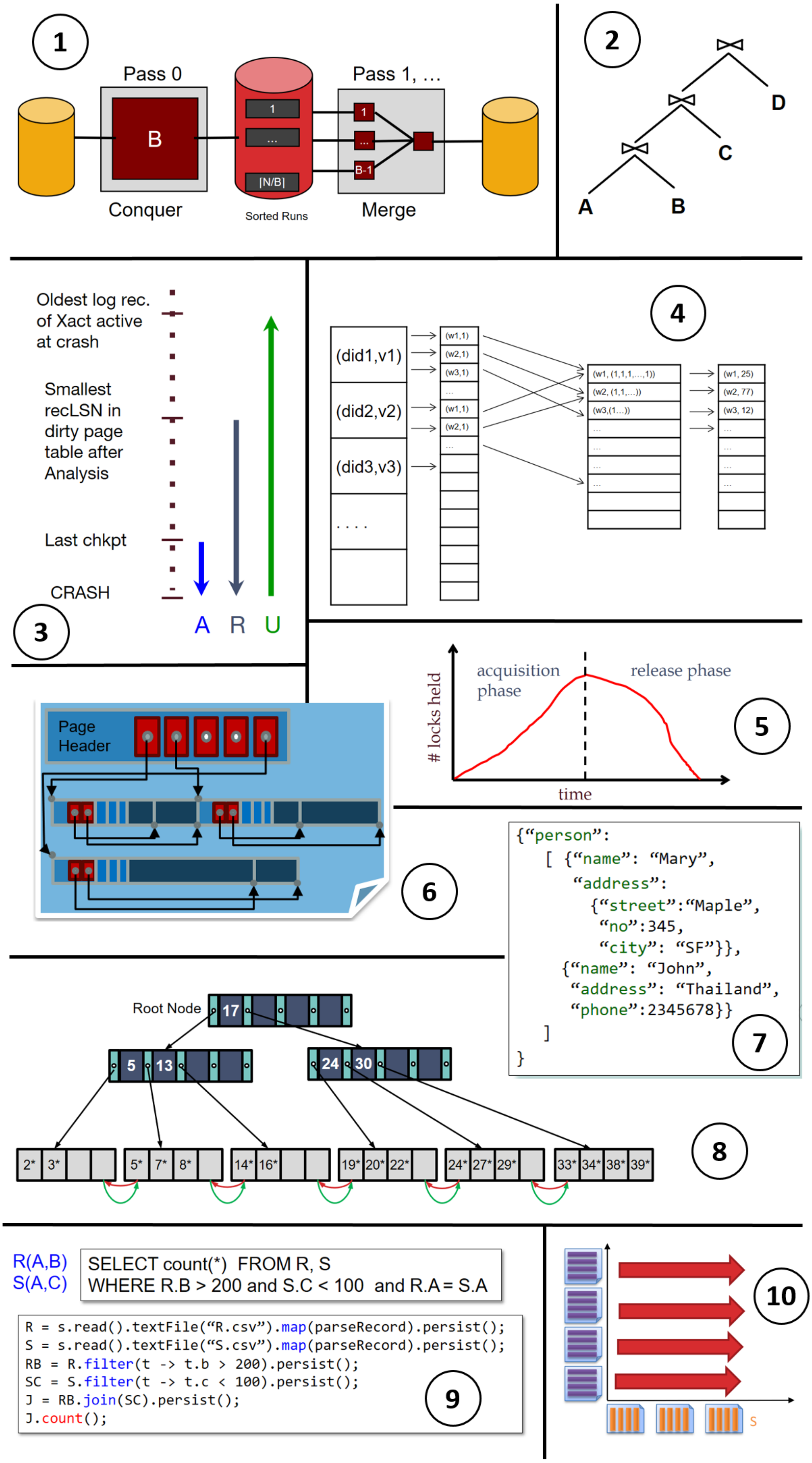
1. **[6 points] Basics**

For each sub-figure in Fig. **??**, select the letter corresponding to the best description.

A. Left Deep Tree
B. Key Compression
C. B+ Tree
D. Spark
E. Nested Loops Join
F. Sort Merge Join
G. Page Nested Loop Join
H. Slotted Page
I. Variable Length Tuple
J. Fixed Length Tuple
K. Buffer Frame
L. Sort-based Group-by
M. Recovery
N. Two Phase Lock
O. JSON
P. ISAM
Q. External Sort
R. MapReduce

Solution

1. Q
2. A
3. M
4. R
5. N
6. H
7. O
8. C
9. D
10. G



```
{"person":
   [ {"name": "Mary",
     "address":
       {"street":"Maple",
        "no":345,
        "city": "SF"}},
    {"name": "John",
     "address": "Thailand",
     "phone":2345678}}
   ]
}
```

SELECT count(*) FROM R, S
WHERE R.B > 200 and S.C < 100 and R.A = S.A

```
R = s.read().textFile("R.csv").map(parseRecord).persist();
S = s.read().textFile("S.csv").map(parseRecord).persist();
RB = R.filter(t -> t.b > 200).persist();
SC = S.filter(t -> t.c < 100).persist();
J = RB.join(SC).persist();
J.count();
```

Figure 1: Basics of Database.

2. **[4 points] SQL**

Which of the following expressions computes the matrix vector product:

$$(\mathbf{A}\mathbf{x})_i = \sum_{k=1}^{d} A_{ik} x_k$$

Assume **A** and **x** have compatible dimensions and there is only one correct answer.

A.

```
SELECT A.row AS row, SUM(A.value * x.value) AS value
FROM A JOIN x
ON A.col = x.row
GROUP BY A.row
```

B.

```
SELECT A.row AS row, SUM(A.value * x.value) AS value
FROM A JOIN x
ON A.row = x.row
GROUP BY A.col
```

C.

```
SELECT x.row AS row, SUM(A.value * x.value) AS value
FROM A JOIN x
ON A.col = x.row
GROUP BY A.col
```

D.

```
SELECT A.row AS row, A.value * x.value AS value
FROM A JOIN x
ON A.row = x.row
```

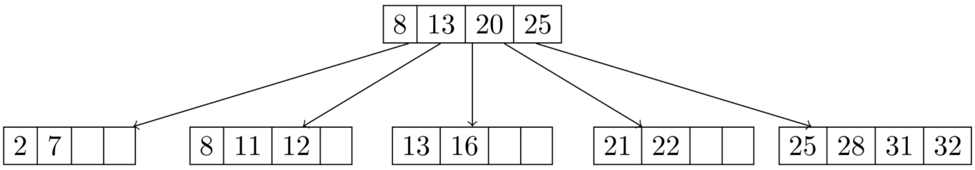
Solution

A.

The final answer should be grouped by the row of A and therefore sum over the product along the columns.

II B+ Trees and Buffer Management [10 points]

1. [**4 points**] **Index and B+ Trees**
   Consider the following B+ tree of order 2.



(a) [**2 points**] How many nodes split when you insert 27?

Solution

- To insert 27, we follow the rightmost pointer in the tree from the root node because $27 > 8$, $27 > 13$, $27 > 20$, $27 > 25$.
- We get to the leaf node containing (25, 28, 31, 32) and attempt to insert 27, but the leaf node is at maximum capacity and the insertion breaks the occupancy invariant. Therefore we must split the leaf into (25, 27) and (28, 32, 21) and copy up key 27 because this is a leaf node.
- We attempt to insert 27 into the parent node containing (8, 13, 20, 25) but the root is also at maximum capacity. We split it into (8, 13) and (25, 28) pushing up key 20 because it is an inner node. Our final tree looks like and we split two nodes.

(b) [**2 points**] After inserting 27 into the tree, you also insert 26. How many nodes split as a result of inserting 26?

Solution

- To insert 26, we first look at the root node and follow the right pointer because $26 \geq 20$. We then follow the middle pointer at the inner node below the root node because $26 > 25$ but $26 < 28$. We get to the leaf node containing (25, 27)
- We attempt to insert 26 into the leaf node containing (25, 27) and since 3 is less than the occupancy invariant 4 we insert 26 and are done. We did not split any nodes.

2. [**6 points**] **Buffer Management**

Supposed we have a buffer pool size of 4 pages, and the following access pattern:
A P P L E S A N D B A N A N A S A N D O R A N G E S
Assume that pages are unpinned immediately (ignore pinning).

(a) [**2 points**] What is the number of cache hits if we use an LRU replacement policy? Assume we are starting from a cold (empty) cache.

Solution
8

(b) [**2 points**] What is the number of cache hits if we use an MRU replacement policy? Assume we are starting from a cold (empty) cache

Solution
5

(c) [**2 points**] What is the number of cache hits if we use a CLOCK replacement policy? Assume we are starting from a cold (empty) cache.

Solution

8

III Sorting and Join Algorithms [10 points]

1. **[4 points] External Sorting**
   Assume that each page is 4 KB large, and that you have a 24KB buffer pool (with 6 frames).

   (a) **[2 points]** How many passes would it take to externally sort an 512KB file? Include the initial sorting pass and subsequent merging passes in your answer. You need to simplify your answer.

   > **Solution**
   > 3: The file contains 128 pages, and B = 6, and thus we need $\left(1 + \left\lceil \log_5 \left(\left\lceil \frac{128}{6} \right\rceil \right) \right\rceil = 3\right)$ passes.

   (b) **[2 points]** What would be the total cost in I/Os for this external sort?

   > **Solution**
   > 768: 2 128 #passes = 256*3 = 768 (I/Os).

2. **[6 points] Join Algorithms**
   Consider a new case, i.e. B¿4 pages worth of buffer space, and relations M and N of size ¿ B. Please fill the blanks below with "always", "sometimes" or "never".

   (a) **[2 points]** Block nested loop join is _______ better than page-oriented nested loop join.

   > **Solution**
   > Always: Block nested loop join is page-oriented on steroids, and steroids are always good.

   (b) **[2 points]** Sort-merge join is _______ better than hash-join.

   > **Solution**
   > Sometimes: Hash join is cooler if one relation is really small and one is very large. Sort dominates if you have lots of duplicate join keys.

   (c) **[2 points]** Hybrid Hash-Join is _______ better than block-nested loops join.

   > **Solution**
   > Sometimes: Nested loops works for non-equijoins and hash does not. For equijoins, hash is often better since it only makes a small number of passes over each relation, whereas block nested-loops still may visit the inner relation many times. If one relation fits in memory, the two algorithms are about equivalent.

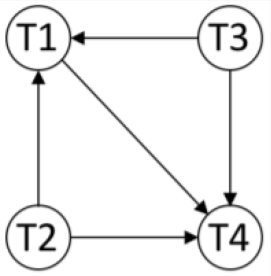
IV Transactions and Concurrency [10 points]

1. **[6 points] Transactions**
Consider the following schedule. (For each of the questions below, you may mark zero (ϕ), one or more than one of the choices.)

	T1	T2	T3	T4
1	R(A)			
2		R(A)		
3			R(C)	
4			W(C)	
5		R(B)		
6		W(B)		
7	R(B)			
8				R(B)
9	R(C)			
10				R(C)
11				W(B)
12		commit		
13			commit	
14	commit			
15				commit

(a) **[2 points]** Please draw the conflict dependency graph of the above schedule?

Solution



(b) **[2 points]** Which of the following schedules below are conflict equivalent to the schedule above?
A. T3, T1, T2, T4
B. T2, T3, T1, T4
C. T4, T3, T1, T2
D. T1, T2, T3, T4

Solution
B
Just by doing a topological sort, we know that T4 has to be last and T1 has to be second to last. Whether T2 or T3 comes first is irrelevant since they have no incoming edges, so we have 2 possible conflict equivalent schedules.

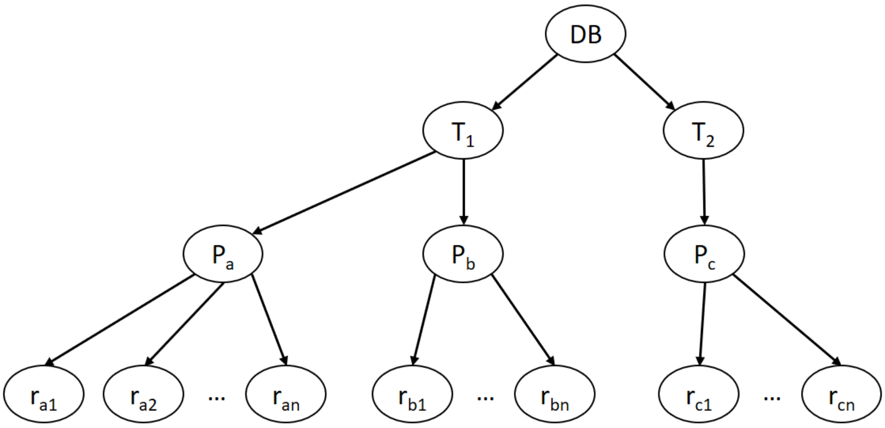
(c) **[2 points]** Select one or more than one true statement(s):
A. In Strict 2PL, we can give up locks after aborting but before rollback is complete.
B. Schedules that are conflict serializable can not produce a cyclic dependency graph.
C. 2PL does not promise conflict serializability.
D. Strict Two Phase Locking ensures that we do not have deadlocks.

Solution
B.

- A is false. For A, you must wait until rollback is complete before giving up locks.

- B is correct. For B, 2PL enforces conflict serializability but may not allow all conflict serializable schedules (e.g. W1(X), R2(X), W1(Y), R2(Y) is impossible under 2PL).
- C is false.
- D is false. For D. Strict 2PL does not help us avoid deadlocks: if T1 wants X(A), X(B), and T2 wants X(B), X(A), we can get into a deadlock if T1 acquires X(A) and T2 acquires X(B).

2. [**4 points**] **Lock Manager**
Given the database system shown below:



(a) [**2 points**] Which lock modes (including S, X, IS, IX, or SIX) on which resources are necessary to read P_a?

(b) [**2 points**] Which lock modes (including S, X, IS, IX, or SIX) held by other transactions on P_a would prevent us from modifying r_{a1}?

Solution

(a) We would need the IS lock mode on DB and T1, and the S lock mode on P_a. This allows us to read from P_a while restricting other transactions as little as possible.

(b) S, SIX, and X lock modes held by other transactions on P_a would prevent us from holding an X lock on r_{a1}, which is necessary to modify r_{a1}. IX and IS locks would not prevent us, as the actual X or S locks held by other transactions are not necessarily on r_{a1}.

V Logging and Recovery [10 points]

1. [**4 points**] **Basics**
   Select the correct choices in the following questions:

   (a) Write Ahead Logging describes a protocol where updated pages must be written to disk before a crash.
   A. True
   B. False

   (b) During a transaction abort, we undo all data updates made by the transaction.
   A. True
   B. False

   (c) In ARIES, UPDATE log records contain no information of the previous state of the page.
   A. True
   B. False

   (d) The recovery manager is responsible for Atomicity and Consistency, as defined by the ACID acronym.
   A. True
   B. False

   (e) If PageLSN is greater than the max LSN flushed so far (flushedLSN), we can safely write this page to disk.
   A. True
   B. False

   (f) In ARIES recovery, after the analysis phase, the recLSN of each page in the dirty page table must be larger than the pageLSN of the corresponding page.
   A. True
   B. False

   (g) When aborting a transaction, it is necessary to modify pages on disk.
   A. True
   B. False

   (h) Write-Ahead Logging (WAL) guarantees that a transactions log records are flushed to disk before the transaction commit.
   A. True
   B. False

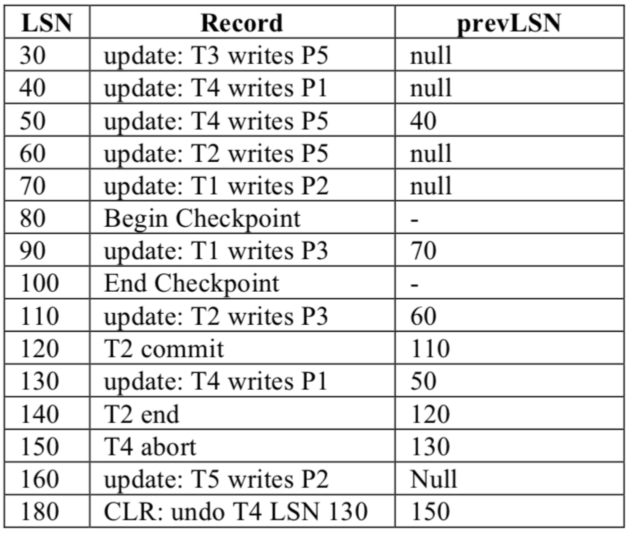
   Solution

   (a) B
   (b) A
   (c) B
   (d) B
   (e) B
   (f) B
   (g) A
   (h) A

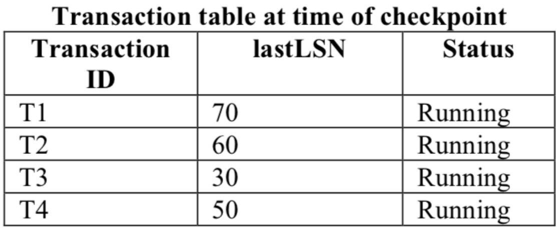
2. [**6 points**] **Recovery**
   Your database server has just crashed due to a power outage. You boot it back up, find the following log and checkpoint information on disk, and begin the recovery process. Assume we use a STEAL/NO FORCE recovery policy.

   (a) [**3 points**] At the end of the Analysis phase, what transactions will be in the transaction table, and with what lastLSN and Status values?

   Solution

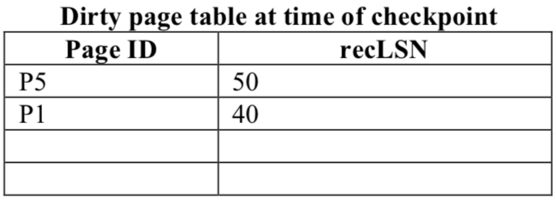


LSN	Record	prevLSN
30	update: T3 writes P5	null
40	update: T4 writes P1	null
50	update: T4 writes P5	40
60	update: T2 writes P5	null
70	update: T1 writes P2	null
80	Begin Checkpoint	-
90	update: T1 writes P3	70
100	End Checkpoint	-
110	update: T2 writes P3	60
120	T2 commit	110
130	update: T4 writes P1	50
140	T2 end	120
150	T4 abort	130
160	update: T5 writes P2	Null
180	CLR: undo T4 LSN 130	150



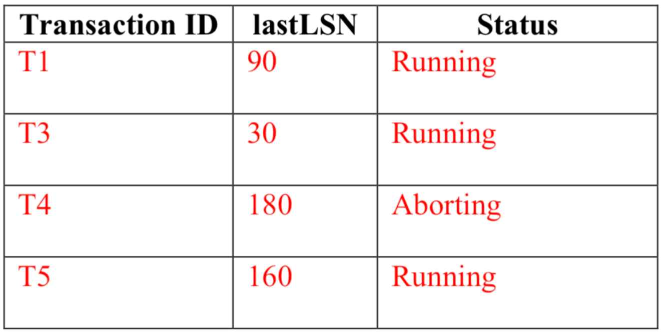
Transaction table at time of checkpoint

Transaction ID	lastLSN	Status
T1	70	Running
T2	60	Running
T3	30	Running
T4	50	Running



Dirty page table at time of checkpoint

Page ID	recLSN
P5	50
P1	40

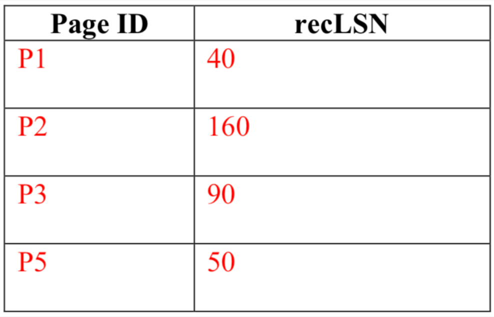


Transaction ID	lastLSN	Status
T1	90	Running
T3	30	Running
T4	180	Aborting
T5	160	Running

We also accept the answer by moving the status of transactions in the above table from "Running" to "Aborting".

(b) [**3 points**] At the end of the Analysis phase, what pages will be in the dirty page table, and with what recLSN values?

Solution



Page ID	recLSN
P1	40
P2	160
P3	90
P5	50

VI Database Design [10 points]

1. [**4 points**] **ER Modeling**
   In this question, we will choose to model elements of a baseball league using ER-Diagrams. We have 4 entities: **Supporter**, **Team**, **Player** and **Agent**. Refer to the Figure below for the following questions.

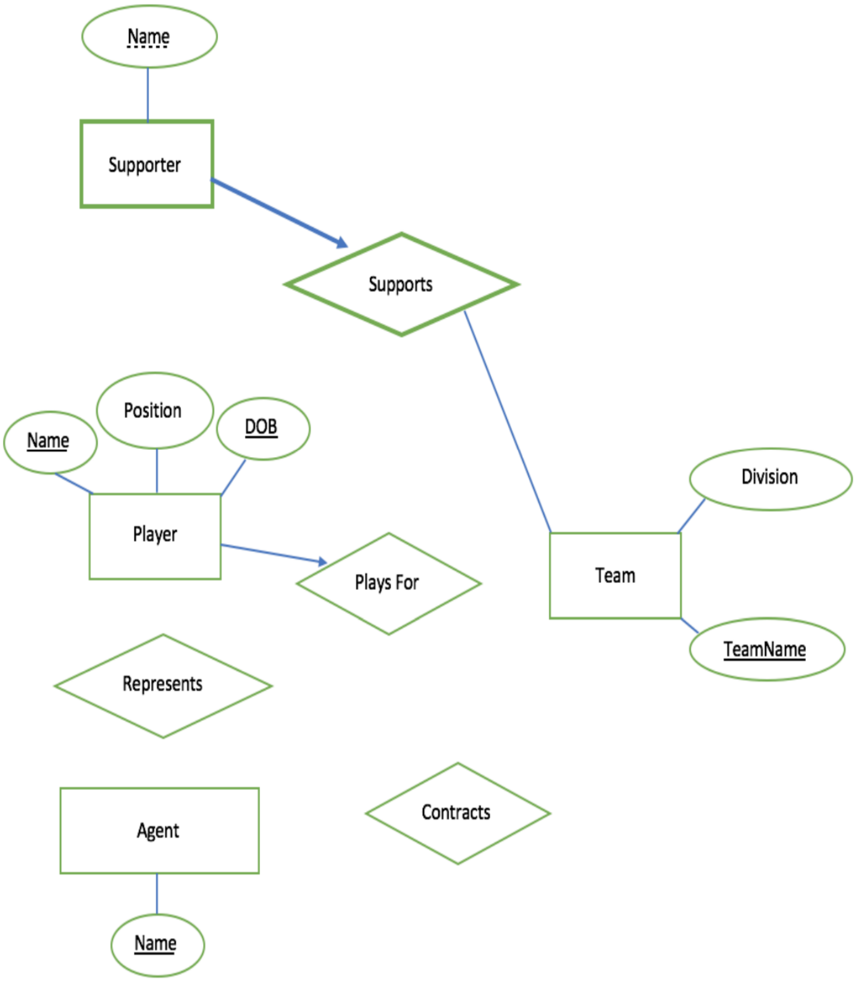


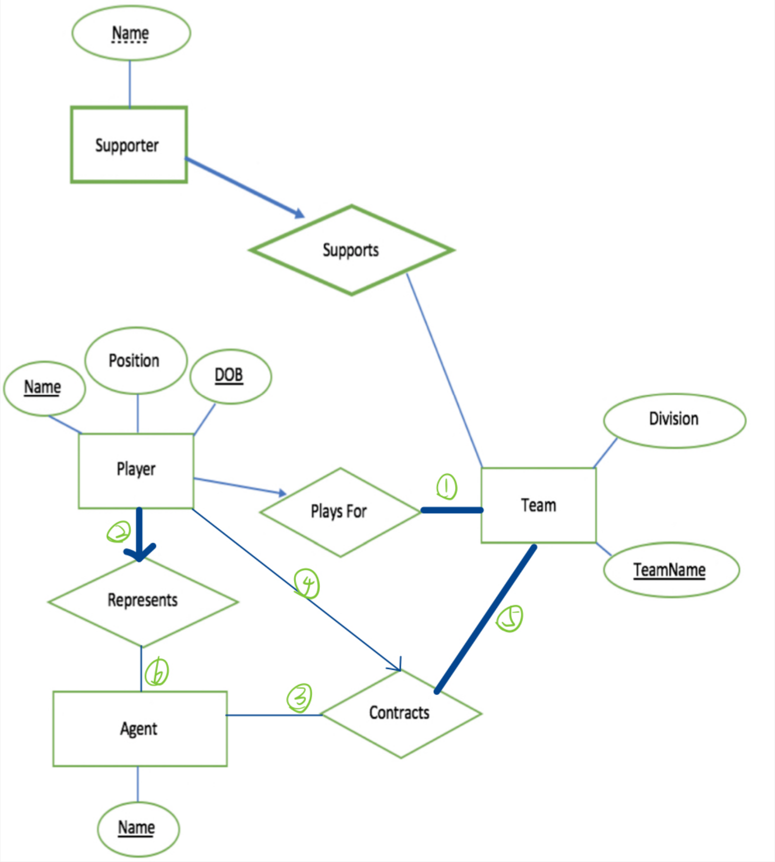
Figure 2: ER model

   (a) [**3 points**] There are some elements of the ER-diagram are missing in Fig. **??**, fill them in so that the ER-diagram satisfies the following constraints.

   - Each team must have at least one player on its roster.
   - Each player must have exactly one agent to represent him/her.
   - Each agent can sign as many contracts as he or she wants to, or none at all.
   - Each player can have at most one contract.
   - A team has at least one contract.
   - Relation **Represents** involves entities **Player** and **Agent**; relation **Contracts** involves entities **Player**, **Agent** and **Team**; relation **Plays For** involves entities **Team** and **Player**

   Hint: you should add 6 edges in total.

   Solution



(b) **[1 point]** As **Supporter** is a weak entity that must have unique and total participation in the **Supports** relation, its primary key is determined by which attribute(s) in the ER-diagram?
A. Name
B. TeamName
C. Name, TeamName
D. None of them

Solution
C

2. **[3 points] Functional Dependencies**
We have a relation R(A, B, C, D, E). We are told that the set of functional dependencies is F = E → BC, A → B, C → D, AD → C.

(a) **[1 point]** What are A+, C+ and E+?

(b) **[1 point]** Is the attribute set ACE the superkey for relation R?

(c) **[1 point]** Is relation R already in Boyce-Codd Normal Form (BCNF)?

Solution

(a) A+: AB, C+: CD, E+: BCDE.

(b) Yes. ACD+: ABCDE.

(c) No. None of the the FDs are a superkey of R and none of them are trivial FDs.

3. **[3 points] Normalization and Decomposition**
Decompose ABCDEFGH into Boyce-Codd Normal Form (BCNF) in the order of the following FDs: G → H; EF → B; EA → C; FH → D.

Solution
At step 1 (G → H), we get ABCDEFG and GH.
At step 2 (EF → B), we get ACDEFG and EFB and GH.
At step 3 (EA → C), we get ADEFG and EAC and EFB and GH.

At step 4 (GF → D), we get AEFG and GFD and EAC and EFB and GH.
(Note: at step 4, GF → D is obtained by combining G → H and FH → D.)

VII Parallel Query Processing [10 points]

1. [**4 points**] **Partition**
   Assume that we have 5 machines and a 1000 page **students(sid, name, gpa)** table. Initially, all of the pages start on one machine. Assume each page is 1KB.

   (a) [**2 points**] How much network cost does it take to round-robin partition the table?

   (b) [**2 points**] How many IOs will it take to execute the following query:

```
SELECT *
FROM students
WHERE name = 'John'
```

Solution

(a) 800 KB.
In round robin partitioning the data is distributed completely evenly. This means that the machine the data starts on will be assigned 1/5 of the pages. This means that 4/5 of the the pages will need to move to different machines, so the answer is: 4/5 * 1000 * 1 = 800 KB.

(b) 1000 IOs.
When the data is round robin partitioned we have no idea what machine(s) the records we need will be on. This means we will have to do full scans on all of the machines so we will have to do a total of 1000 IOs.

2. [**6 points**] **Parallel Algorithm**
   Assume that we have a 100 page table R that we will join with a 400 page table S. All the pages start on machine 1, and there are 4 machines in total with 30 buffer pages each.

   (a) [**2 points**] How many passes are needed to do a parallel unoptimized Sort-Merge Join (SMJ) on the data? For this question a pass is defined as a full pass over either table (so if we have to do 1 pass over R and 1 pass over S it is 2 total passes).

   (b) [**2 points**] How many passes are needed to a parallel Grace Hash Join on the data?

   (c) [**2 points**] We want to calculate the max in parallel using hierarchical aggregation. What value should each machine calculate and what should the coordinator do with those values?

Solution

(a) 7 passes.
2 passes to partition the data (1 for each table). Each machine will then have 25 pages for R and 100 pages for S. R can be sorted in 1 pass but S requires 2 passes. Then it will take 2 passes to merge the tables together (1 for each table). This gives us a total of 2 (partition) + 1 (sort R) + 2 (sort S) + 2 (merge relations) = 7 passes.

(b) 4 passes.
We will need 2 passes to partition the data across the four machines and each machine will have 25 pages for R and 100 for S. We don't need any partitioning passes because R fits in memory on every machine, so we only need to do build and probe, which will take 2 passes (1 for each table). This gives us a total of 4 passes.

(c) Each machine should calculate the max and the coordinator will take the max of those maxes.

VIII Distributed Transactions and NoSQL [10 points]

1. [**3 points**] **2-Phase Commit**
   Select the correct choices in the following questions:

   (a) If we are recovering and see a PREPARE record, it means we must have sent out a YES vote.
   A. True
   B. False

   (b) If the coordinator is recovering and sees a COMMIT record, we can locally commit and end the process there.
   A. True
   B. False

   (c) Suppose a participant receives a prepare message for a transaction and replies VOTE-YES. Suppose that they were running a wound-wait deadlock avoidance policy, and a transaction comes in with higher priority. The new transaction will abort the prepared transaction.
   A. True
   B. False

   Solution

   (a) B, False.
   We could have crashed between logging PREPARE and sending VOTE-YES.

   (b) B, False.
   There is no guarantee that all participants received a COMMIT record, so we must rerun phase 2 first.

   (c) B, False.
   Transactions that are prepared must be ready to commit if the coordinator tells them to, so they cannot be aborted by anyone other than the coordinator.

2. [**2 points**] **OLTP and OLAP**
   Are the following workloads better characterized as Online Transaction Processing (OLTP) or Online Analytical Processing (OLAP)? Select the correct choices in the following questions:

   (a) Placing orders and buying items in an online marketplace.
   A. OLTP
   B. OLAP

   (b) Analyzing trends in purchasing habits in an online marketplace.
   A. OLTP
   B. OLAP

   Solution

   (a) A, OLTP.
   OLTP workloads involve high numbers of transactions executed by many different users. Queries in these workloads involve simple lookups more often than complex joins. In this case, when a user buys an item, the site might perform actions like looking up the item and updating its quantity, recording a new purchase, etc.

   (b) B, OLAP.
   OLAP workloads involve read-only queries and typically include lots of joins and aggregations. Often, workloads executed for analysis and decision making are OLAP workloads.

3. [**5 points**] **NoSQL**
   Select the correct choices in the following questions:

(a) As methods for scaling dataset, partitioning is effective for write-heavy workloads, and replication is effective for read-heavy workloads.
A. True
B. False

(b) Partitioning and replication are often used together in real systems to leverage the performance benefits and to make the system more fault-tolerant.
A. True
B. False

(c) The consistency in the distributed systems is the same concept with the consistency found in ACID (Transactions).
A. True
B. False

(d) JSON and XML are referred as semi-structured data formats, and can express complex data structures, which may be arranged into nested structures.
A. True
B. False

(e) JSON is self-describing, and it is designed for efficient storage and retrieval from disk.
A. True
B. False

Solution

(a) A, True.

(b) A, True.

(c) B, False.

(d) A, True.

(e) B, False.

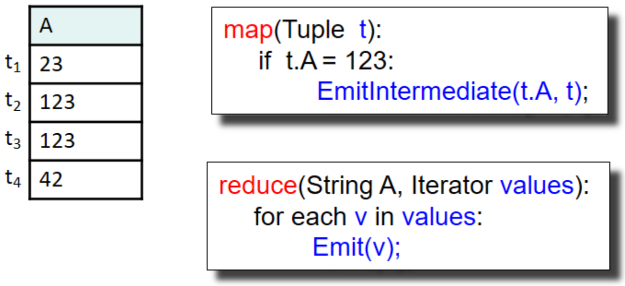
IX MapReduce and Spark [10 points]

1. **[3 points] Basics**
Select all the true statement(s):
A. In MapReduce, the Map phase applies a function in parallel to every element of a set of data, and the Reduce phase combines the results of the Map phase into the desired data output.
B. In Distributed File System (DFS), a file system in which large files are partitioned into smaller files, and then distributed and replicated several times on different nodes for fault tolerance.
C. Both the Map and Reduce phases can be split among different workers on different machines, with workers performing independent tasks in parallel.
D. Since the Reduce phase must wait until the entire Map phase completes, we must restart the entire Map phase if one map task fails.
E. Resilient Distributed Datasets (RDDs) are not written to disk in intermediate steps but are rather stored in main memory.
F. Similar with MapReduce, Spark needs to write intermediate results to disk.

Solution
A, B, C, E

2. **[7 points] MapReduce**

(a) **[2 points]** Fig. **??** shows a relation $R(A)$ with the MapReduce implementation of the Selection operator $\sigma_{A=123}(R)$.



	A
t_1	23
t_2	123
t_3	123
t_4	42

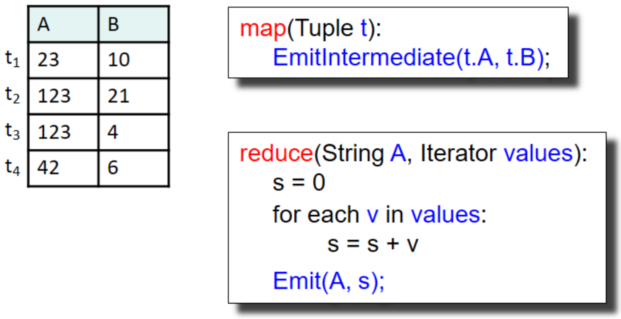
```
map(Tuple  t):
   if  t.A = 123:
          EmitIntermediate(t.A, t);
```

```
reduce(String A, Iterator values):
   for each v in values:
          Emit(v);
```

Figure 3: Selection

(1) What is the output of the Map function?
(2) What is the output of the Reduce function?

(b) **[2 points]** Fig. **??** shows a relation $R(A, B)$ with the MapReduce implementation of the Group-By operator $\gamma_{A=123,\mathrm{SUM}(B)}(R)$.



	A	B
t_1	23	10
t_2	123	21
t_3	123	4
t_4	42	6

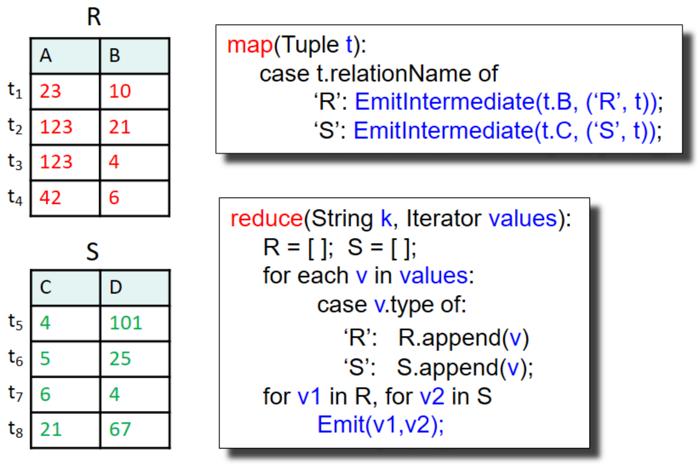
```
map(Tuple t):
   EmitIntermediate(t.A, t.B);
```

```
reduce(String A, Iterator values):
   s = 0
   for each v in values:
          s = s + v
   Emit(A, s);
```

Figure 4: Group-By

(1) What is the output of the Map function?
(2) What is the output of the Reduce function?

(c) **[3 points]** Fig. **??** shows two relations $R(A, B)$ and $S(C, D)$ with the MapReduce implementation of the Hash-Join operator $R(A, B) \bowtie_{B=C} S(C, D)$.



R

	A	B
t1	23	10
t2	123	21
t3	123	4
t4	42	6

S

	C	D
t5	4	101
t6	5	25
t7	6	4
t8	21	67

```
map(Tuple t):
    case t.relationName of
        'R': EmitIntermediate(t.B, ('R', t));
        'S': EmitIntermediate(t.C, ('S', t));
```

```
reduce(String k, Iterator values):
    R = [ ];  S = [ ];
    for each v in values:
        case v.type of:
            'R':   R.append(v)
            'S':   S.append(v);
    for v1 in R, for v2 in S
        Emit(v1,v2);
```

Figure 5: Hash-Join

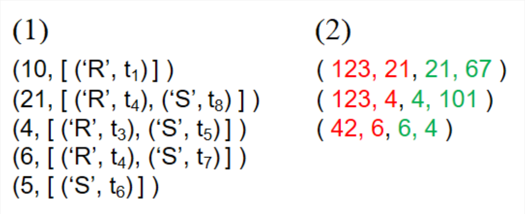
(1) What is the output of the Map function?

(2) What is the output of the Reduce function?

Solution

(a) (1) $(123, [t_2, t_3])$
(It is also correct: $(123, t_2)$ and $(123, t_3)$)
(2) (t_2, t_3).

(b) (1) $(23, [t_1])$, $(42, [t_4])$, $(123, [t_2, t_3])$
(It is also correct: $(23, t_1)$, $(42, t_4)$, $(123, t_2)$, $(123, t_3)$)
(2) $(23, 10)$, $(42, 6)$, $(123, 25)$.



(1)

(10, [('R', t_1)])
(21, [('R', t_4), ('S', t_8)])
(4, [('R', t_3), ('S', t_5)])
(6, [('R', t_4), ('S', t_7)])
(5, [('S', t_6)])

(2)

(123, 21, 21, 67)
(123, 4, 4, 101)
(42, 6, 6, 4)

Figure 6: Answer of (c). For (1), it is also correct if the results are not shuffled based on the intermediate key.

(c)

X Data Mining and Machine Learning [10 points]

1. **[2 points] Basics**
Select all the true statement(s):
A. The order of the basic KDD (Knowledge Discovery in Database) process is Data Selection, Data Cleaning, Data Mining & ML, Evaluation.
B. The ultimate goal in machine learning is to find a model that best fits the training data.
C. The k-means algorithm is guaranteed to converge to the global optimum.
D. A common form of feature engineering on continuous data is one-hot-encoding.

Solution
A

2. **[4 points] K-Means**

(a) **[2 points]** Which of the following can act as possible termination conditions in k-means? (you may select one or more than one of the choices.)
A. For a fixed number of iterations.
B. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.
C. Centroids do not change between successive iterations.
D. Terminate when the objective value falls below a threshold.

Solution
A, B, C, D.
All four conditions can be used as possible termination condition in K-Means clustering:
This condition limits the runtime of the clustering algorithm, but in some cases the quality of the clustering will be poor because of an insufficient number of iterations. Except for cases with a bad local minimum, this produces a good clustering, but runtimes may be unacceptably long. This also ensures that the algorithm has converged at the minima. Terminate when the objective value falls below a threshold. This criterion ensures that the clustering is of a desired quality after termination. Practically, it's a good practice to combine it with a bound on the number of iterations to guarantee termination.

(b) **[2 points]** In which of the following cases will k-means clustering fail to give good results? (you may mark zero (ϕ), one or more than one of the choices.)
A. Data points with outliers.
B. Data points with different densities.
C. Data points with round shapes.
D. Data points with non-convex shapes.

Solution
A, B, D.
K-Means clustering algorithm fails to give good results when the data contains outliers, the density spread of data points across the data space is different and the data points follow non-convex shapes.

3. **[4 points] Linear Regression**
Given n independent and identically distributed (i.i.d.) training samples $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, with the i-th sample $x_i, y_i \in \mathbb{R}$, $i = 1, 2, ..., n$. Consider the linear model:

$$y = x\theta + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

where $\mathcal{N}(0, \sigma^2)$ denotes the Guassian distribution with mean 0 and variance σ^2. Assume that the training data has been centralized, such that the intercept can be ignored in the above linear model. Obviously, the probability density function of y conditioned on x and θ follows the Gaussian distribution $\mathcal{N}(x\theta, \sigma)$, which is defined by

$$p(y|x, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(y - x\theta)^2). \tag{2}$$

The model parameter θ can be estimated based on Maximum Likelihood Estimation (MLE), which aims to maximize the likelihood function:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} P(x_i, y_i|\theta) \\ &= \prod_{i=1}^{n} P(y_i|\theta, x_i)P(x_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y_i - x_i\theta)^2}{2\sigma^2})P(x_i). \end{aligned} \tag{3}$$

Hint: $P(x_i)$ $(i = 1, 2, ..., n)$ can be treated as a constant in the above formulation.

(a) **[2 points]** Show the log-likelihood function $\log L(\theta)$. (The base of the log function is e.)

(b) **[2 points]** Apply MLE to calculate the closed-form solution of θ.

Solution

(a) According to the definition of the likelihood function, we have

$$\begin{aligned} \log L(\theta) &= \sum_{i=1}^{n} -\frac{1}{2}\log 2\pi - \log\sigma - \frac{(y_i - x_i\theta)^2}{2\sigma} + \log P(x_i) \\ &= -\frac{1}{2\sigma}\sum_{i=1}^{n}(y_i - x_i\theta)^2 + C, \end{aligned} \tag{4}$$

where C denotes a constant, and it is irrelevant to θ.

(b) Based on (a), we have

$$\hat{\theta} = \underset{\theta}{\text{argmax}} \log L(\theta) = \text{argmin} \sum_{i=1}^{n}(y_i - x_i\theta)^2. \tag{5}$$

By seting the derivative of above objective function w.r.t. θ equal to 0, we have

$$\theta = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}. \tag{6}$$