DATA 514 Section 5 Worksheet

Name: _____

Lecture Take-aways

1. Decompose Restaurants(id, name, rating, popularity, rec?) on id->name, rating. On which relations do the FDs remain?

2. Why is full table scan a slow way to search for matching rows?

3. Do indexes make these queries run faster, slower, or the same?

	Faster	Slower	Same
`SELECT * FROM Flights WHERE fid = 1234;`			
`SELECT * FROM Flights WHERE date < '2000/01/01';`			
`SELECT * FROM Flights WHERE dest = 'NYC';`			
`INSERT INTO Flights VALUES (3243, '2022/04/20', 'DL', 'SEA','NYC',159);`			

4. Does a (date, origin) index make this query run faster, slower, or the same?

	Faster	Slower	Same
`SELECT * FROM Flights WHERE origin = 'SEA';`			

5. When might an index make a SELECT-FROM query run *slower?*

6. What indexes could we make on Users?

`SELECT * FROM Users, Assets WHERE Users.id = Assets.uid`	

`SELECT *`
`  FROM Users`
`WHERE Users.score > 95`	

`SELECT *`
`  FROM Users`
`WHERE Users.age > 21`	

7. What datasets could you join against? What join key would you use?

Section Practice

Normalization

Question 2 (4 points)

Given R(A, B, C, D, E), and functional dependencies: A → C, BD → A, D → E

1. **Find the following closures: {A}+, {B}+, {D}+, and {BD}+**

2. **Decompose R into BCNF. In each step, explain which functional dependency you used to decompose and explain why further decomposition is needed. Your answer should consist of a list of table names and attributes. Make sure you indicate the keys for each relation.**

Indices

Question 2 (5 points)

Given the following:

```
CREATE TABLE Items(
id                        INTEGER PRIMARY KEY,
name                      VARCHAR(64) NOT NULL,
quantity                  VARCHAR(64)  -- quantity that we currently
-- have in stock. May be 0,
-- but cannot be negative
);

CREATE TABLE Orders (
id                        INTEGER PRIMARY KEY,
customerName              VARCHAR(64) NOT NULL,
address                   VARCHAR(64),
     orderDate            DATE  -- assume this type is equivalent to
-- INTEGER; ie >, =, MAX(), etc are
                              -- well defined.
);

CREATE TABLE OrderItems (
oid                       INTEGER REFERENCES Orders,
iid                       INTEGER REFERENCES Items,
isFulfilled               INTEGER -- boolean
quantity                  INTEGER -- guarenteed > 0
);
```

Consider the following query load. Is it a point selection or a range scan based query? What type of index is optimal, clustered or unclustered?

Part A (1 point)

This query is executed several times per second:

```
-- Identify the items we need to put into a shipment SELECT
oi.iid, SUM(oi.quantity)
FROM Orders o, OrderItems oi WHERE
o.id = oi.oid
AND oi.isFulfilled = 0;
```

Part B (1 point)

This query is executed once per hour:

```
-- Identify the out-of-stock items that are preventing old orders
-- from being fulfilled
SELECT i.id, i.name, SUM(oi.quantity) FROM
Orders o, OrderItems oi, Items i WHERE o.id
= oi.oid AND oi.iid = i.id AND
oi.isFulfilled = 0
AND i.quantity = 0 AND
o.orderDate < ? GROUP BY
i.id, i.name ORDER BY
o.orderDate;
```

Part C (1 point)

These queries are executed approximately once every second:

```
-- Record a new order
INSERT INFO Orders VALUES (?, ?, ?);

-- Record their ordered items. An "average" order has 3-5 items,
but
-- it is possible to have a single-item order. Orders are capped
to 256
-- unique times, enforced by the database application. INSERT
INTO OrderItem VALUES (?, ?, 0, ?);
INSERT INTO OrderItem VALUES (?, ?, 0, ?); INSERT
INTO OrderITem VALUES (?, ?, 0, ?);
-- ...etc...
```

Part D (1 point)

This query is executed many many times per second, but only during business hours:

```
-- Ship an item UPDATE
OrderItems SET isFulfilled
= 1 WHERE oid = ?
AND iid = ?;
```

Part E (1 point)

Lastly, approximately 10% of queries are none of these; they're a mixture of tools that aren't run frequently or adhoc reports run by analysts looking for interesting patterns (maybe brand loyalty to Bose headphones?). We don't want to optimize our database for these queries.

Data Governance

Question 3 (6 points)

Your friend, the one that created the multi-billion dollar startup Kelp, has a new idea: Tuber Eats, a tuber ordering and delivery platform for everything from potatoes to yams.

- **They've hired you to work on internal analytics, which generates reports based on the data contained in every order:**
  - **IP address**
  - **Restaurant**
  - **Purchase date, time, amount, items**
  - **User name, precise age, gender**
- **However, you're only required to generate these three reports:**
  - **Item popularity by area and month (eg, "jicama sales in Seattle are higest in June")**
  - **Item popularity by age and gender (eg, "retired women purchase taro twice as often as a median user")**
  - **Restaurant popularity by time (eg, "Red Radish is most popular 11.30-11.40 and 12.30-12.40")**

For each piece of information in the order, describe how (or whether) you would store it, ie if anonymization and what type, what duration of storage, etc.

Part A (1 point)

IP address, assume that an IP address is able to locate a user's neighborhood within a city.

Part B (1 point)

Restaurant

Part C (1 point)

Purchase Timing, the date and time

Part D (1 point)

Purchase Amount

Part E (1 point)

Purchased Items

Part F (1 point)

Precise Age

ETL

Question 4 (4 points)

Part A (1 point)

Using Indeed, builtin, or other jobsearch site, find a job posting requiring ETL skills. Write out the company and the job description requirement for ETL (just the bullet point or line).

Part B (1 point)

Go to the company's website. Look at the products and/or services that they offer.

Suggest types of data that the company may be working with and potential relationships between the data, a simplified or non formal ER diagram.

Part C (2 points)

What do you think the company's weakness is? Extracting, Transforming, or Loading? Why? (2-3 sentence explanation)

Part D (Bonus)

How would you construct a resume or cover letter with this in mind?
