CS490 Exam 1
(Rao)

Name: ______________________

September 25, 2017

Instructions

1. Please write your name on the sheet.
2. This is a closed book exam and should be completed within 75 minutes.
3. This exam has 4 questions. You are required to answer all the questions for full credit.
4. Read each question carefully. Your answers should be precise, complete, and correct to receive full credit. If your sentences are poorly constructed and we do not understand your answer, then some points will be deducted.
5. If you do not understand a question, please ask but do not try to discuss your answers.
6. All the best!

Question 1 - Chord (10 points)

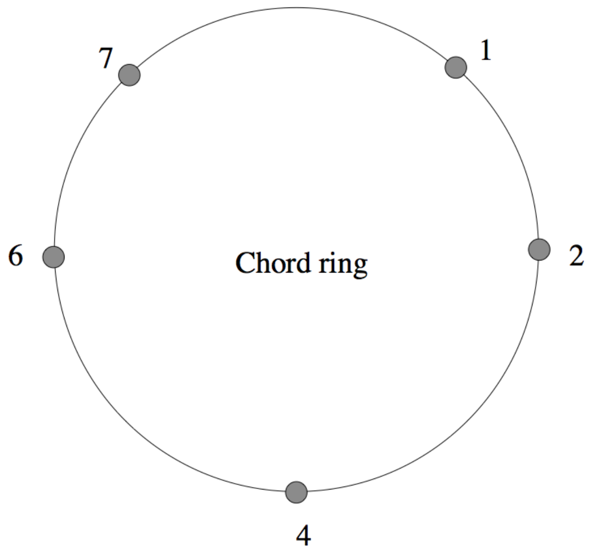


Figure 1: Chord system with 5 peers

Suppose a Chord DHT uses 3-bit identifier space for peers and keys. As shown in Figure 1, there are five peers mapped to Chord ids 1, 2, 4, 6, and 7. Answer the following questions. (Do not worry about which keys are actually present in the system.)

1. Write down the finger tables at peers 6, 7, and 2. (6 points)

2. Suppose a lookup for key 3 is issued at peer 6. Explain how it is found using the finger tables. (4 points)

Question 2 - HBase (10 points)

Consider the logical table shown below with two column families 'User' and 'Location'.

Rowkey	User		Location	
	Name	EmpID	Timezone	Country
`www.John.com`	John	1001	us-east	USA
`www.Mary.com`	Mary	2001	fr	France

1. Write down the list of HBase commands to construct this table. (7 points)

2. Write the HBase command to output the attributes Name, Timezone, and Country for rowkey `www.John.com`. (3 points)

Question 3 - Hive (10 points)

Write down the set of Hive QL commands to do the following operations.

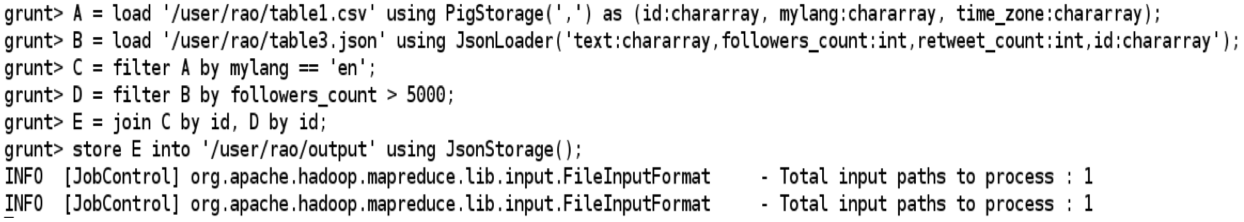
1. Create a table `tweet_info` with attributes `tweet_id` (INT), `tweet_text` (STRING), and `tweet_language` (STRING). (3 points)

2. Create a table `user_info` with attributes `user_name` (STRING), `tweet_id` (INT), and `followers_count` (INT). (3 points)

3. Assuming `tweet_info` and `user_info` contain data, join these tables on `tweet_id` and output the attributes `tweet_id`, `tweet_text`, `tweet_language`, `user_name`, and `followers_count`. (4 points)

Question 4 - Pig (6 points)

In Figure 2, a set of Pig commands have been posed by a user to achieve a specific data processing task. Explain the purpose of each Pig command.



```
grunt> A = load '/user/rao/table1.csv' using PigStorage(',') as (id:chararray, mylang:chararray, time_zone:chararray);
grunt> B = load '/user/rao/table3.json' using JsonLoader('text:chararray,followers_count:int,retweet_count:int,id:chararray');
grunt> C = filter A by mylang == 'en';
grunt> D = filter B by followers_count > 5000;
grunt> E = join C by id, D by id;
grunt> store E into '/user/rao/output' using JsonStorage();
INFO  [JobControl] org.apache.hadoop.mapreduce.lib.input.FileInputFormat     - Total input paths to process : 1
INFO  [JobControl] org.apache.hadoop.mapreduce.lib.input.FileInputFormat     - Total input paths to process : 1
```

Figure 2: Example

THIS PAGE IS THE LAST PAGE AND IS INTENTIONALLY LEFT BLANK FOR SCRATCH WORK.