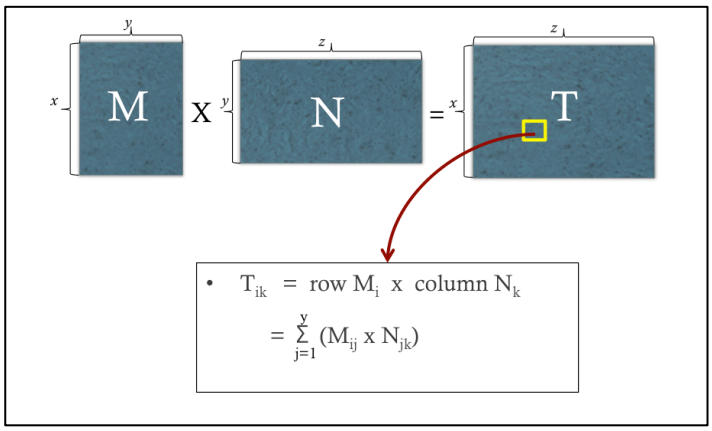
CS585: Big Data Management
Final Exam

(11 Questions, each is worth 5 Points)

Student Name:

Q1: Draw a figure showing the differences between shared-nothing, shared-memory, and shared-disk architectures.

Q2: The figure below shows the matrix multiplication problem covered in class. Assume each matrix (M and N) is stored in a separate file (with the same matrix name). Each line in the files is in the format (*i, j, v*) where "i" is the row number, "j" is the column number, and "v" is the cell value. Assume we want to solve the problem using Hadoop.



(a) **[1 Point]** How many mapreduce jobs are needed?
(b) **[4 Points]** Describe the map input, map output, reduce input, reduce output of the job(s).

Q3: In Spark, what is the difference between Action operations and Transformation operations? Give two examples of each.

Q4: In Spark, draw a figure that shows the difference between a Narrow Dependency and a Wide Dependency? And Then, Give one example operation in each case.

Q5: Given the following operations/algorithms:
K-Means Clustering, Logistic Regression, Aggregation operation, Selection operation, Building the model of Naïve Bayes Classifier, Page Rank

State which one(s) will have similar performance (and also mention why) if executed on either Hadoop or Spark infrastructures. Assume the input data in all cases is initially stored on HDFS.

Q6: Specify the data model of each of the following databases?

Q7: In NoSQL Databases, we studied two concepts CAP & BASE.

(a) [3 Points] The following figure shows the CAP theorem. Put the following databases on the triangle edge matching with the DB characteristics.

Relational DB, MongoDB, HBase, Cassandra, CouchDB

(b) [2 Points] The keyword "BASE" is an abbreviation of what?

Q8: MongoDB has the concept of a "ReplicaSet", which is a group of machines controlled by one primary plus multiple secondaries. What does each of the following properties mean? And give a scenario in which such property setting is useful?

(a) Priority = 0
(b) Hidden = True
(c) SlaveDelay = 100

Q9: **(a) [2 Points]** Give two reasons on why Hadoop and Spark are not considered as database engines unlike RDBMS, MongoDB, HBase?

(b) [3 Points] Which of the following systems rely on Zoopkeeper? And state why it does (or does not) relay on Zookeeper.

Relational DB, MongoDB, HBase, Cassandra, CouchDB

Q10: In Apache Hive, answer the following questions:

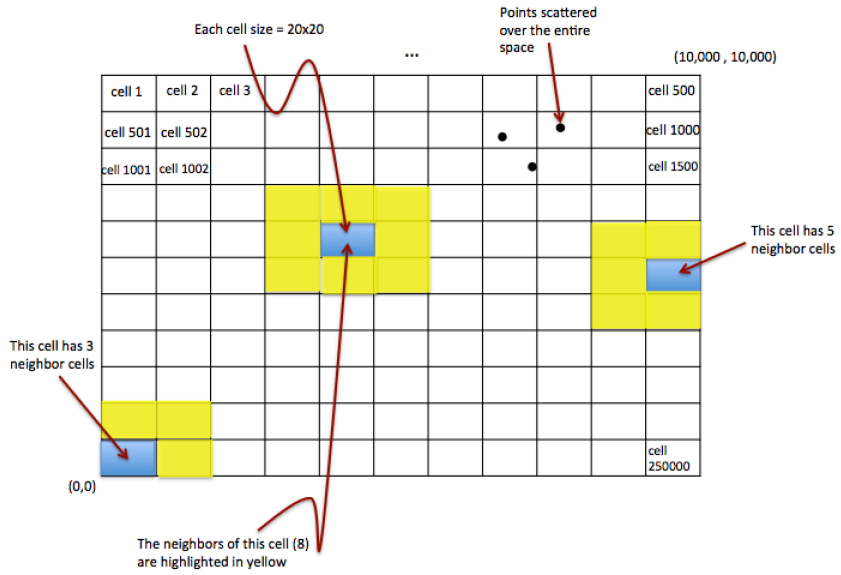
(a) [3 Points] The following Hive statement creates a partitioned table "R". Draw a figure that shows how R will be stored in HDFS (that is, the structure that hive will maintain to store R)

```
Create Table R (a int, b int, c string) Partitioned By
(country string, date string);
```

(b) [2 Points] Apache Hive has a unique component in its architecture that does not exist in Apache Pig. What is this component and why it exists in Hive alone?

Q11: [This problem is given to you in HW3 (Spark Project)]

Assume a two-dimensional space that extends from 1…10,000 in each dimension as shown in the figure below. There are points scattered all around the space. The space is divided into pre-defined grid-cells, each of size 20x20. That is, there is 500,000 grid cell in the space. Each cell has a unique ID as indicated in the Figure. Given an ID of a grid cell, you can calculate the row and the column it belongs to using a simple mathematical equation.



Neighbor Definition: For a given grid cell X, N(X) is the set of all neighbor cells of X, which are the cells with which X has a common edge or corner. The Figure illustrates different examples of neighbors. Each non-boundary grid cell has 8 neighbors. However, boundary cells will have less number of neighbors (See the figure). Since the grid cell size is fixed, the IDs of the neighbor cells of a given cell can be computed using a formula (mathematical equations) in a short procedure.

Example: N(Cell 1) = {Cell 2, Cell 501, Cell 502}

N(Cell 1002) = {Cell 501, Cell 502, Cell 503, Cell 1001, Cell 1003, Cell 1501, Cell 1502, Cell 1503}

Relative-Density Index: For a given grid cell X, I(X) is a decimal number that indicates the relative density of cell X compared to its neighbors. It is calculated as follows.

I(X) = X.count / Average (Y_1.count, Y_2.count, …Y_n.count)

Where "X.count" means the count of points inside grid cell X, and {Y_1, Y_2, …, Y_n} are the neighbors of X. That is N(X) = {Y_1, Y_2, …, Y_n}. If "Average (Y_1.count, Y_2.count, …Y_n.count)" = 0, then I(X) should be set to zero.

The question is: Assume we are using Spark RDDs. The first initial RDD has lines, where each line is one point in the space represented as (x,y). As discussed in class, draw a diagram showing the sequence of RDDs that need to be generated to find the TOP 50 I(X) grid cells in a scalable way. For each transformation (E.g., from RDD_i to RDD_{i+1}) indicate the output of both the map and the reduce functions (assuming each transformation is done in the map-reduce style).

Left Blank for Answers

Left Blank for Answers

Left Blank for Answers