Homework 6 | Parallelized Data Management

Updates made to the assignment spec after release are highlighted in red.

Objectives: To understand how a dataset's underlying distribution and partitioning algorithm work together to produce skew. To practice decomposing relational queries into MapReduce format.

Due date: Thursday, May 22 @ 9:00pm

Median completion time (24sp): **4.5 hours**

Resources

- Pen and paper, or any drawing tools you prefer (e.g., PowerPoint, draw.io).

Problem Set

1. Hopefully a happy little bonus for you to start with :)

2. You have a table *R* containing 50M rows and wish to partition it into 2 <= k <= 32 shards. You decide to partition on attribute *A*, whose range is *documented* as [-2^{31}, 2^{31}]. You need to choose from partitioning schemes (uniform, range, or hash). (Hint: Where do the values fall on a number line? How many unique values are there, and are they distributed?)

   Recall that in lecture, we defined *skew* as the existence of a partition R_i such that $|R_i| >> |R|/k$. For the purposes of this question, we will add specificity to this definition by requiring $|R_i| > 1.5 * |R|/k$.

   a. Although *A*'s range is declared as [-2^{31}, 2^{31}], in actuality *A* uniformly takes on values from [0, 2^{31}]. Which partitioning schemes would **not** yield skew for all possible values of *k*?
   b. In actuality *A* uniformly takes on values that are powers of 2 *and also* in the range [-2^{2}, 2^{31}]. Thus, *A* only consists of [-2^{2}, -2^{1}, -2^{0}, 2^{0}, 2^{1},... 2^{31}]. Which partitioning schemes would **not** yield skew for all possible values of *k*?

c. In actuality *A* uniformly takes on values that are multiples of *k* in its declared range [-2^{31}, 2^{31}]. Thus, *A* contains values such as {$-2k$, $173k$, $-2^{16}k$, etc}. Which partitioning schemes would **not** yield skew for all possible values of *k*?

d. In actuality *A* uniformly takes on values that are *power-of-2 multiples* of *k* in its declared range. Thus, *A* only consists of {$-2^{(31-\log k)}k$, … $-2^{1}k$, $-2^{0}k$, $2^{0}k$, $2^{1}k$, ... $2^{(31-\log k)}k$}. Which partitioning schemes would **not** yield skew for all possible values of *k*?

3. Consider our relations `Payroll(UserID, Name, Job, Salary)` and `Regist(UserID, Car, Year)` and the following query:

```
WITH OldestCar AS
     (  SELECT P1.Job AS Job, MIN(R1.Year) AS Year
          FROM Payroll AS P1, Regist AS R1
         WHERE P1.UserID = R1.UserID
      GROUP BY P1.Job)
SELECT P.Name, OC.Year
FROM Payroll AS P,
     Regist AS R,
     OldestCar AS OC
WHERE P.UserID = R.UserID
  AND P.Job = OC.Job
  AND R.Year = OC.Year
  AND R.Year <= 2020
```

Use FJWGHOS to construct the *unoptimized relational algebra tree*; you do not need to submit this tree. Next, optimize your tree by pushing the selection ($\sigma_{\text{R.Year <= 2020}}$) as far down as possible and possibly reordering the joins. Submit your *optimized* tree.

4. Implement the previous question's query as a series of MapReduce tasks; place your selection ($\sigma_{\text{R.Year <= 2020}}$) in the location suggested by your *optimized* tree.

Hints: Your third MapReduction will take the output of the previous two MRs as input.

5. Now, consider your *unoptimized* tree from question 3; if you had used this tree to implement your MR, in which task would you have implemented the selection ($\sigma_{\text{R.Year <= 2020}}$)?

6. Next, please reflect on your own personal experience. Specifically:

a. What is one thing that you *learned* while doing this assignment?
b. What is one thing that *surprised you* while doing this assignment?
c. What is one *question that you still have* after doing this assignment?

7. Lastly, please answer the feedback questions:

a. How many hours did it take you to finish this assignment, including time to set up your computer (if necessary)?

b. How many of those hours did you feel were valuable and/or contributed to your learning?
c. Did you collaborate with other students on this assignment? If so, approximately how many people did you collaborate with? Do not include yourself or course staff in the count.

Submission Instructions

You should submit the questions on Gradescope under HW6.