Exercise 1 (Star Join) 0 points

We consider a data structure that often arises in commercial data. A large company wants to keep track of all their sales. To do that they keep a fact table $F = (A_1, \dots, A_m)$, where every tuple represents a single sale. The key attributes A_i represent the important components of the sale, such as purchaser ID, the ID of the purchased item, the store etc. For every key item A_i, there is a dimension table $D_i = (A_i, B_i)$ which contains further information, like address of the purchaser or prize of the item. Usually the size f of the fact table is much larger than the sizes d_i of the dimension tables or the number ob key attributes. Analytic queries on these data bases usually join the fact table with several of the dimension tables and aggregate the result. Such a join is called *star join*.
Consider the following Map-Reduce algorithm for the star join $F \bowtie D_1 \bowtie \dots \bowtie D_m$. Let s be the number of reducers, let $s_1, \dots, s_m \in \mathbb{N}$ such that $\prod_{i=1}^m s_i = s$ be the shares of the key attributes and let $h_1, \dots, h_m$ be independently chosen truly random hash functions $h_i \colon V_i \to [s_i]$, where V_i is the set of values of the attribute A_i.

MAP:	On input $(F, (a_1, \dots, a_m))$,	emit $((h_1(a_1), \dots, h_m(a_m)), (a_1, \dots, a_m))$.
	On input $(D_i, (a_i, b_1))$,	emit all key-value pairs $((p_1, \dots, p_{i-1}, h_i(a_i), p_{i+1}, \dots, p_m), (a_i, b_1))$ for $p_j \in [s_j]$ for all $j \neq i$.
REDUCE:	On input $(\bar{p}, \mathit{values})$,	compute $\mathcal{Q}(\bar{p}) := \mathcal{F}(\bar{p}) \bowtie \mathcal{D}_1(\bar{p}) \bowtie \dots \bowtie \mathcal{D}_m(\bar{p})$, where $\mathcal{F}(\bar{p}) := \{t \mid (F, t) \in \mathit{values}\}$ and $\mathcal{D}_i(\bar{p}) := \{t \mid (D_i, t) \in \mathit{values}\}$, and emit all pairs (Q, t) for $t \in \mathcal{Q}(\bar{p})$.

a) Compute the replication rate of this algorithm.

b) Compute the expected load of this algorithm.

c) Assume $s = \prod_{i=1}^n d_i$. Which choice of the s_i minimizes the values in (a) and (b)?

Hint: You may use the following inequality, known as AM-GM inequality, without proving it. Let $x_1, \dots, x_n \in \mathbb{R}^{\geq 0}$, then

$$\frac{\sum_{i=1}^n x_i}{n} \geq \sqrt[n]{\prod_{i=1}^n x_i}, \tag{1}$$

where equality holds if and only if $x_1 = x_2 = \dots = x_n$.

Solution:

a) The total number of tuples that the map tasks output is

$$f + \sum_{i \in [m]} d_i \prod_{j \neq i} s_j = f + s \sum_{i \in [m]} \frac{d_i}{s_i}.$$

The input size is $f + \sum_{i\in[m]} d_i$, thus the replication rate is

$$\frac{f}{f + \sum_{i\in[m]} d_i} + \frac{s}{f + \sum_{i\in[m]} d_i} \sum_{i\in[m]} \frac{d_i}{s_i}.$$

b) Consider the reducer for the key $\bar{p} = (p_1, \ldots, p_k)$. The load of this reducer is

$$|\mathcal{F}(\bar{p})| + \sum_{i\in[m]} |\mathcal{D}_i(\bar{p})|.$$

Let $i \in [\ell]$ and $t \in F$, for $j \in [m]$ let a_j be the A_j-value of t. Then

$$\begin{aligned}\Pr_{h_1,\ldots,h_k}\left(t \in \mathcal{R}_i(\bar{p})\right) &= \Pr_{h_1,\ldots,h_k}\left(h_j(a_j) = p_j \text{ for all } j \in [m]\right)\\ &= \prod_{j\in[m]} \Pr_{h_j}\left(h_j(a_j) = p_j\right)\\ &= \prod_{j\in[m]} \frac{1}{s_j} = \frac{1}{s}.\end{aligned}$$

Thus

$$\mathrm{E}\left(\left|\mathcal{F}(\bar{p})\right|\right) = \sum_{t\in F} \Pr_{h_1,\ldots,h_k}\left(t \in \mathcal{F}(\bar{p})\right) = \frac{f}{s}.$$

Let $i \in [\ell]$ and $t \in D_i$, let a_i be the A_i-value of t. Then

$$\begin{aligned}\Pr_{h_1,\ldots,h_k}\left(t \in \mathcal{D}_i(\bar{p})\right) &= \Pr_{h_1,\ldots,h_k}\left(h_i(a_i) = p_i\right)\\ &= \Pr_{h_i}\left(h_i(a_i) = p_i\right)\\ &= \frac{1}{s_i}.\end{aligned}$$

Thus

$$\mathrm{E}\left(\left|\mathcal{D}_i(\bar{p})\right|\right) = \sum_{t\in D_i} \Pr_{h_1,\ldots,h_k}\left(t \in \mathcal{D}_i(\bar{p})\right) = \frac{d_i}{s_i}.$$

Thus we get that the expected load is

$$\frac{f}{s} + \sum_{i\in[m]} \frac{d_i}{s_i}.$$

c) Both values are minimized when $\sum_{i\in[m]} \frac{d_i}{s_i}$ is minimized.

For $i \in [m]$, we introduce new variables $x_i = \frac{d_i}{s_i}$ and we rewrite the optimization problem:

Minimize $\sum_{i\in[m]} x_i$ subject to the constraint $\prod_{i\in[m]} x_i = \frac{\prod_{i\in[m]} d_i}{\prod_{i\in[m]} s_i} = 1$.

From (1) we know that $\sum_{i\in[m]} x_i \geq m \sqrt[m]{\prod_{i=1}^{m} x_i} = m$ and the minimum is realized if and only if $x_1 = x_2 = \ldots = x_m$. Thus we get $x_1 = x_2 = \ldots = x_m = 1$ which implies $s_i = d_i$, for all $i \in [m]$.

Exercise 2 (Random Elements from Stream) **8 points**

Suppose you receive a stream $a_1, \ldots, a_n$ of positive integers a_i for all $i \in [n]$. Let $z = a_1 + a_2 + \ldots a_n$ be the sum of all elements in the stream.
Describe a streaming sampling algorithm that has the probability $\frac{a_i}{z}$ to pick the element with the index i and prove its correctness. The algorithm shall use at most $O(log\ z)$ bits of space.
Note: Just like the length n of the stream, the value z is not known to your algorithm in advance.

Solution:

Consider the following algorithm:

- $z \leftarrow 0$
- while there is another element a_{i+1} in the stream do
- $\quad z \leftarrow z + a_i$
- $\quad sample \leftarrow a_i$ with probability $\frac{a_i}{z}$
- return sample

We claim the index of the element returned by the algorithm is i with probability $\frac{a_i}{z}$. For the proof let z_j be the value of z after the j-th iteration of the loop. We show the claim by induction on j.

I.B.: Obviously, after one single step, the probability that our sample is a_1 is exactly $\frac{a_1}{z_1} = 1$.
I.H.: The index of the element *sample* of the algorithm after j many steps is i with probability $\frac{a_i}{z_j}$.
I.S.: Let x_{j+1} be the value of the element *sample* after iteration $j+1$. We have $\Pr(x_{j+1} = a_{j+1}) = \frac{a_{j+1}}{z}$ by line 6 of the algorithm.
Moreover, for some $i < j+1$, we have the following:

$$\begin{aligned}
\Pr(x_{j+1} = a_i) &= \Pr(x_j = a_i) \cdot \left(1 - \frac{a_{j+1}}{z_{j+1}}\right) \\
&\overset{\text{I.H.}}{=} \frac{a_i}{z_j} \cdot \left(1 - \frac{a_{j+1}}{z_{j+1}}\right) \\
&= \frac{a_i}{z_j} \cdot \left(\frac{z_{j+1}}{z_{j+1}} - \frac{a_{j+1}}{z_{j+1}}\right) \\
&= \frac{a_i}{z_j} \cdot \left(\frac{z_{j+1} - a_{j+1}}{z_{j+1}}\right) \\
&= \frac{a_i}{z_j} \cdot \frac{z_j}{z_{j+1}} \\
&= \frac{a_i}{z_{j+1}}
\end{aligned}$$

Exercise 3 (Strongly-2-Universal Hashing) **8+4=12 points**

a) Consider the family

$$\mathcal{H} \coloneqq \{(ax+b) \mod p \colon 0 \leq a, b \leq p-1\}$$

of hash function from $\mathbb{U} = \{0, \ldots, p-1\}$ to $\mathbb{T} = \{0, \ldots, p-1\}$ where p is a prime number. Show that $\mathcal{H}$ is strongly 2-universal.

- Show this directly. In particular, do **not** just apply (the more general) statement from the lecture.
- Point out where you use that p is a prime number.

b) Assume that $|\mathbb{U}| \geq 2$ and let $\mathbb{T}$ be arbitrary (non-empty). Does there exist a strongly 2-universal family $\mathcal{H}'$ of hash function from $\mathbb{U}$ to $\mathbb{T}$ such that $|\mathcal{H}'| < |\mathbb{T}|^2$?

If yes, describe a construction of such a family. If no, prove that no such family exists.

Solution:

a) Note that $|\mathbb{T}| = |\mathbb{U}| \geq 2$. Let $x_1, x_2 \in \mathbb{U}$ be distinct elements, let $y_1, y_2 \in \mathbb{T}$, and let $h_{a,b}$ be the function that maps x to $(ax+b) \mod p$. Following the definition of strongly k-universal families (Definition 8.10), we need to show that

$$\Pr_{h \in \mathcal{H}} \big(h(x_1) = y_1 \wedge h(x_2) = y_2\big) = \tfrac{1}{p^2}.$$

Note that if $(a,b) \neq (a',b')$, then $h_{a,b} \neq h_{a',b'}$. (If $a = a'$ but $b \neq b'$, then $a \cdot 0 + b \not\equiv a' \cdot 0 + b' \pmod p$. If $a \neq a'$ and $b = b'$ then $a \cdot 1 + b \not\equiv a' \cdot 1 + b' \pmod p$.)

For $h = h_{a,b}$, the event „$h(x_1) = y_1 \wedge h(x_2) = y_2$“ gives a system of linear equations over $\mathbb{F}_p$:

$$\begin{aligned} ax_1 + b &= y_1 \\ ax_2 + b &= y_2 \end{aligned}$$

with $\cdot$ and $+$ being multiplication and addition in $\mathbb{F}_p$. This system has exactly one solution

$$\begin{aligned} a &= \tfrac{y_2 - y_1}{x_2 - x_1} \\ b &= y_1 - ax_1 \end{aligned}$$

in $\mathbb{F}_p$, as p is prime. That is, for all distinct $x_1, x_2 \in \mathbb{U}$, and all $y_1, y_2 \in \mathbb{T}$, there is a unique pair $(a,b) \in \{0, \ldots, p-1\}^2$ with $h_{a,b}(x_1) = y_1$ and $h_{a,b}(x_2) = y_2$. As drawing h from $\mathcal{H}$ uniformly at random is equivalent to drawing (a,b) from $\{0, \ldots, p-1\}$ uniformly at random, it follows that

$$\Pr_{h \in \mathcal{H}} \big(h(x_1) = y_1 \wedge h(x_2) = y_2\big) = \tfrac{1}{p^2}.$$

b) No, there is no such family. Let $x_1, x_2 \in \mathbb{U}$ be two fixed distinct elements. Suppose $\mathcal{H}'$ is strongly 2-universal. In particular, $\Pr_{h \in \mathcal{H}}\left(h(x_1) = y_1 \wedge h(x_2) = y_2\right) = \frac{1}{|\mathbb{T}|^2} > 0$ for all $(y_1, y_2) \in \mathbb{T}^2$. Thus, for all $(y_1, y_2) \in \mathbb{T}^2$ there exists at least one function $h \in \mathcal{H}'$ such that $h(x_1) = y_1$ and $h(x_2) = y_2$. Clearly, these functions are different for different pairs (y_1, y_2). Thus $|\mathcal{H}'| \geq |\mathbb{T}|^2$.