Algorithms for Provisioning Queries and Analytics

Sepehr Assadi
University of Pennsylvania

Joint work with Sanjeev Khanna (Penn), Yang Li (Penn) and Val Tannen (Penn)
Motivating Example

A Database:

<table>
<thead>
<tr>
<th>Sales venue</th>
<th>Commission</th>
<th>Reputation</th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Venue 1</td>
<td>15%</td>
<td>200</td>
<td>20k</td>
</tr>
<tr>
<td>Venue 2</td>
<td>20%</td>
<td>300</td>
<td>30k</td>
</tr>
<tr>
<td>Venue 3</td>
<td>0%</td>
<td>900</td>
<td>50k</td>
</tr>
<tr>
<td>Venue 4</td>
<td>10%</td>
<td>700</td>
<td>10k</td>
</tr>
</tbody>
</table>
Motivating Example

A Database:

<table>
<thead>
<tr>
<th>Sales venue</th>
<th>Commission</th>
<th>Reputation</th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Venue 1</td>
<td>15%</td>
<td>200</td>
<td>20k</td>
</tr>
<tr>
<td>Venue 2</td>
<td>20%</td>
<td>300</td>
<td>30k</td>
</tr>
<tr>
<td>Venue 3</td>
<td>0%</td>
<td>900</td>
<td>50k</td>
</tr>
<tr>
<td>Venue 4</td>
<td>10%</td>
<td>700</td>
<td>10k</td>
</tr>
</tbody>
</table>

Analyst:
- What is the correlation between commission and reputation of a sale venue with its revenue?
  - E.g. using ordinary least square (OLS) method, i.e., l2-regression.
Motivating Example

A Database:

<table>
<thead>
<tr>
<th>Sales venue</th>
<th>Commission</th>
<th>Reputation</th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Venue 1</td>
<td>15%</td>
<td>200</td>
<td>20k</td>
</tr>
<tr>
<td>Venue 2</td>
<td>20%</td>
<td>300</td>
<td>30k</td>
</tr>
<tr>
<td>Venue 3</td>
<td>0%</td>
<td>900</td>
<td>50k</td>
</tr>
<tr>
<td>Venue 4</td>
<td>10%</td>
<td>700</td>
<td>10k</td>
</tr>
</tbody>
</table>

Analyst:

- What is the correlation between commission and reputation of a sale venue with its revenue under different scenarios?
  - E.g. using ordinary least square (OLS) method, i.e., l2-regression.
Motivating Example

- **Hypothetical:** an event *retaining* a subset of entries.
  - What if we only consider Tablets?
  - What if we only consider US products?
  - What if we only consider US retailers?

- **Scenario:** a collection of hypotheticals.
  - What if we consider Tablets and US products?
  - What if we consider Tablets but not US retailers?
Motivating Example

- **Hypothetical**: an event *retaining* a subset of entries.
  - What if we only consider Tablets?
  - What if we only consider US products?
  - What if we only consider US retailers?

- **Scenario**: a collection of hypotheticals.
  - What if we consider Tablets and US products?
  - What if we consider Tablets but not US retailers?
Motivating Example

Original database + identified hypotheticals

<table>
<thead>
<tr>
<th>Sales venue</th>
<th>Commission</th>
<th>Reputation</th>
<th>Revenue</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Venue 1</td>
<td>15%</td>
<td>200</td>
<td>20k</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venue 2</td>
<td>20%</td>
<td>300</td>
<td>30k</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venue 3</td>
<td>0%</td>
<td>900</td>
<td>50k</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venue 4</td>
<td>10%</td>
<td>700</td>
<td>10k</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Motivating Example

### Database of Scenario H1

<table>
<thead>
<tr>
<th>Sales venue</th>
<th>Commission</th>
<th>Reputation</th>
<th>Revenue</th>
<th>H1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Venue 1</td>
<td>15%</td>
<td>200</td>
<td>20k</td>
<td></td>
</tr>
<tr>
<td>Venue 2</td>
<td>20%</td>
<td>300</td>
<td>30k</td>
<td></td>
</tr>
<tr>
<td>Venue 4</td>
<td>10%</td>
<td>700</td>
<td>10k</td>
<td></td>
</tr>
</tbody>
</table>
### Motivating Example

**Database of Scenario H2 + H3**

<table>
<thead>
<tr>
<th>Sales venue</th>
<th>Commission</th>
<th>Reputation</th>
<th>Revenue</th>
<th>H2</th>
<th>H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Venue 1</td>
<td>15%</td>
<td>200</td>
<td>20k</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venue 3</td>
<td>0%</td>
<td>900</td>
<td>50k</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venue 4</td>
<td>10%</td>
<td>700</td>
<td>10k</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Motivating Example

**Question:** Can we help the analyst in answering (possibly all) scenarios?
Motivating Example

**Question:** Can we help the analyst in answering (possibly all) scenarios?

Possible solutions:
1. Give the database to the analyst.
2. Give the answers to the analyst.
Motivating Example

**Question:** Can we help the analyst in answering (possibly all) scenarios?

Possible solutions:
1. Give the database to the analyst.
2. Give the answers to the analyst.

How bad can it be?
Consider a database of 1TB size and $k=50$ hypotheticals
Motivating Example

**Question:** Can we help the analyst in answering (possibly all) scenarios?

Possible solutions:
1. Give the database to the analyst.
2. Give the answers to the analyst.

How bad can it be?
Consider a database of 1TB size and $k=50$ hypotheticals
1. = Give the analyst 1TB of data (+ lots of re-computation).
2. = Give the analyst $2^{50}$ bit = 128TB of data!
Provisioning

**Question:** Can we summarize the database w.r.t to the hypotheticals into a small-space representation (a provisioned sketch) such that every scenario can be answered using only the sketch?

Small-space: *polynomial* in $k$ and $\log(n)$, e.g. $O(k^2\log(n))$. 

**Provisioning**

**Question:** Can we summarize the database w.r.t to the hypotheticals into a small-space representation (a provisioned sketch) such that every scenario can be answered using only the sketch?

Small-space: polynomial in $k$ and $\log(n)$, e.g. $O(k^2\log(n))$. 

query + hypotheticals  

compression phase  

extraction phase  

a scenario  

answer to the scenario
Queries considered

1. **COUNT()**: number of the entries.
2. **SUM()**: summation of the entries (over a column).
3. **MEDIAN()**: median of the entries (over a column).
4. **LIN_REG()**: linear regression over the entries.
5. **Boolean queries** followed by numerical queries.
6. **GROUPBY** and **HAVING** clauses.
Results

Numerical queries:
COUNT(), SUM(), MEDIAN(), LIN_REG():

- Bad news:
  - They admit no non-trivial provisioned sketch.
  - I.e., any sketch for provisioning requires $\min(2^k, n)$ space.
Results

Numerical queries: COUNT(), SUM(), MEDIAN*(()), LIN_REG():

- **Bad news:**
  - They admit no non-trivial provisioned sketch.
  - I.e., any sketch for provisioning requires min(2^k, n) space.

- **Good news:**
  - They admit poly(k, log(n), 1/ε)-size provisioned sketch for 1±ε approximation answer.
Results

Other results:

- **UCQ queries** followed by numerical queries can be compactly provisioned.
- **UCQ queries with negation** cannot be compactly provisioned.
- **Datalog** queries cannot be compactly provisioned.
- **GROUPBY** clauses of "small" size can be compactly provisioned.
- **HAVING** clauses cannot be compactly provisioned.
**COUNT() query**

**Thm.** There is a compact provisioned sketch for $(1 \pm \epsilon)$-approximation of the \textsc{ COUNT()} query.

- size $O(k^2/\epsilon^2 \log \log(n))$. 
Thm. There is a compact provisioned sketch for \((1\pm \varepsilon)\)-approximation of the \texttt{COUNT()}
query.
- size \(O(k^2/\varepsilon^2 \log \log(n))\).

How good is this?
Consider a database of 1TB size and \(k=50\) hypotheticals and assume \(\varepsilon = 1\%\), i.e., we allow 1% error.
Thm. There is a compact provisioned sketch for \((1\pm\varepsilon)\)-approximation of the \texttt{COUNT()} query.

- size \(O(k^2/\varepsilon^2 \log \log(n))\).

How good is this?
Consider a database of 1TB size and \(k=50\) hypotheticals and assume \(\varepsilon = 1\%\), i.e., we allow 1\% error.

- \(n \leq 2^{43}\) hence \(\log \log(n) \leq 6\).
- The constant in \(O(-)\) is less than 100.
- The total space is less than 60MB.
Provisioning \texttt{COUNT()} (simplified version)

Compression:
1. Pick a random hash function $h:[n]$ to $[n]$ and let $t = 96/\varepsilon^2$.
2. For each hypothetical $H_i$, store the $t$ smallest value of $h(x \cdot \text{ID})$ for $x$ in $H_i$.

Extraction:
1. Given the scenario $S$, let $r$ be the $t$-th smallest number in the stored numbers for hypotheticals in $S$.
2. Return $t \times n/r$ as the answer.
Lower bounds

Provisioning setting:

- Query + hypotheticals
- Linear time algorithm
- Compression phase
- Polynomial time algorithm
- Extraction phase
- Poly(log(n),k)-size sketch
- a scenario
- Answer to the scenario
Lower bounds

Lower bound setting:

query + hypotheticals → Computationally unbounded → Compression phase → Extraction phase → Computationally unbounded → a scenario → answer to the scenario
Lower bounds

Lower bound setting:

- **Compression phase**: Computationally unbounded
- **Extraction phase**: Computationally unbounded

The scenario is known after $o(n)$ entries can be seen.

- **Query + hypotheticals**: a scenario
- **Answer to the scenario**:
Lower bounds

Lower bound setting:

\[ \min(2^k, n) \]-size sketch is needed!

Computationally unbounded

Compression phase

Extraction phase

\( o(n) \) entries can be seen after scenario is known

query + hypotheticals

a scenario

answer to the scenario
Coverage Problem

A canonical hard problem for provisioning:

Given \( k \) subsets \( H_1, H_2, \ldots, H_k \) of \([n]\), compress them to answer the following scenarios.

Given a subset \( S \) of \([k]\): is \([n]\) covered by the union of \( H_i \) for \( i \) in \( S \)?
Coverage Problem

A **canonical hard** problem for provisioning:

Given $k$ subsets $H_1, H_2, \ldots, H_k$ of $[n]$, compress them to answer the following scenarios.

Given a subset $S$ of $[k]$: is $[n]$ covered by the union of $H_i$ for $i \in S$?

**Thm.** The *coverage* problem requires $\min(n, 2^k)$ size sketches.
Coverage Problem

A **canonical hard** problem for provisioning:

Given $k$ subsets $H_1, H_2, \ldots, H_k$ of $[n]$, compress them to answer the following scenarios.

Given a subset $S$ of $[k]$: is $[n]$ covered by the union of $H_i$ for $i$ in $S$?

**Thm.** The *coverage* problem requires $\min(n, 2^k)$ size sketches.

**Thm.** Exact provisioning schemes for $\text{COUNT}(\cdot)$, $\text{SUM}(\cdot)$, $\text{LIN\_REG}(\cdot)$ can solve *coverage* problem also.
Thank you