Approximate Query Processing: Overview and Challenges

Peter J. Haas

College of Information and Computer Sciences
University of Massachusetts Amherst

Thanks to:
Andrew McGregor
Barzan Mozafari
Approximate Query Processing (APQ)

Data stream → Data (static or dynamic) → Query

- Exact answer (slow)
- Approx. answer (fast)

Data stream → Synopsis → Query

- sample or sketch
AQP is More Important Than Ever

How to deal with data explosion?

Parallel/Distributed Computing

AQP

Can combine both

- Costly for ordinary people (EC2 = $)
- Not eco-friendly
- Some algorithms not embarrassingly parallel
- Concurrent queries degrade performance

- Cheap
- Green

Source: Patrick Cheesman 2016

Source: InsideBIGDATA 2017
APQ Canonical Examples I

**Histogram:**
- `SELECT COUNT(x) WHERE 5.1 < x < 10.3`
- Exact answer: 21
- Approximate answer: 
  \[(4.9/5) \times 21 + (0.3/5) \times 13 = 21.36\]
Sample:
- SELECT SUM(prod) FROM clicks GROUP BY prod
**Sketch**
- SELECT COUNT(DISTINCT x)
- Exact answer: 4
- Approximate answer: \((2/0.413) - 1 = 3.84\)
## A Taxonomy of APQ Problems

<table>
<thead>
<tr>
<th>Static queries</th>
<th>Simple analytics</th>
<th>Complex analytics</th>
<th>Machine Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static queries</td>
<td>Heavy hitters,</td>
<td>Sketches (FM, AMS, LSH, ...)</td>
<td>Spanner (distances)</td>
</tr>
<tr>
<td></td>
<td>Max/min,</td>
<td>Random projections,</td>
<td>Graph mining,</td>
</tr>
<tr>
<td></td>
<td>Quantiles,</td>
<td>Bayesian models</td>
<td>Fixed analytic workflows</td>
</tr>
<tr>
<td></td>
<td>Distinct values,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Frequency moments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predict. queries and data</td>
<td>SPJ+agg queries, L_p distances, Range sums, K-nearest neighbors, Subset sums</td>
<td>Stratified/VarOpt/Measure-biased/CR samples, Sample + index, Workload-based wavelets and histograms SQL queries, Visual analytics, Analytic workflows</td>
<td>Bayesian and maxEnt models, ML workflow</td>
</tr>
<tr>
<td>Ad hoc queries</td>
<td>SPJ+agg queries, Visual analytics</td>
<td>Uniform samples, Multi-dim. histograms Bayesian models SQL queries</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Injected distinct samplers (Quickr)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ad hoc ML</td>
</tr>
</tbody>
</table>

SPJ = Select, Project, Join
**Challenge: Industrial Strength APQ Systems (Mozafari 2017)**

<table>
<thead>
<tr>
<th>OLAP Workloads</th>
<th>TPC-H</th>
<th>TPC-DS</th>
<th>Facebook</th>
<th>Conviva Inc.</th>
<th>Customer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupported Queries</td>
<td>See paper</td>
<td>Full outer joins</td>
<td>Joins of multiple fact tables</td>
<td>Joins of multiple fact tables</td>
<td>Multiple fact joins, nested, textual filters</td>
</tr>
<tr>
<td>Percentage of Supported Queries</td>
<td>68%</td>
<td>&gt; 90%</td>
<td>&gt; 96 %</td>
<td>91%</td>
<td>74%</td>
</tr>
<tr>
<td>Speedup</td>
<td>10x</td>
<td>2x</td>
<td>?</td>
<td>10-200x</td>
<td>2-20x</td>
</tr>
</tbody>
</table>

Source: Mozafari 2017

So far: relatively simple SQL queries
Challenge: Industrial Strength APQ Systems (Mozafari 2017)

Compatibility with existing engines: Middleware required
- Efficiency challenges
- Automatic query rewrite needed

Dealing with existing interfaces
- Compatibility and user friendliness
- High-level accuracy contracts
  (at least p% accurate with p% prob and exist w. p% prob)
Challenge: Industrial Strength APQ Systems (Mozafari 2017)

Query planning

- Different query-plan criteria from traditional query optimization
  - Minimize time to acceptable error or error within time constraint
  - Error can be hard to predict and control
    - So far: Analytical formulas, Bayesian modeling, analytical/Poisson bootstrap
    - A priori error guarantees (sample+seek w. measure-biased sampling, indexes...)
  - Latency is very hard to predict (esp. in parallel/distributed setting)

- Automatically choosing the right synopsis
  - Run a competing set of synopses and combine answers
  - Theory? E.g, space complexity analysis [Kaushik et al. 2005]

- Learning based on prior results + exploration (extend to dynamic data)
Challenge: Industrial Strength APQ Systems

Handling Complex analytics

- Arbitrary SQL aggregate queries
  - Subqueries: [Joshi and Jemaine 2009; Rusu et al. 2015]
  - Quickr [Kandula et al. 2016] inject distinct-samplers into query plan (multiple passes)

- Set-valued queries [Ioannidis and Poosala 1999]

- Modern queries
  - Graph queries
  - ML (coreSets, model management, sampleClean)

- Sequences of analytical operations: error propagation? [Ioannidis & Christodolakis 1991]

- Error estimation and guarantees
  - Even in “simple” SPJ+Agg setting with GROUP-BY and selection predicates
Challenge: APQ for Visual Analytics I

Achieving high interactivity

- Combine ad-hoc sampling with precomputed samples and indexes (e.g., AQUA, BlinkDB, IDEA, VisTrees)
- Reuse results between queries (IDEA, Verdict)
- Predict user behavior to fetch or precompute synopsis of interest (DICE, ForeCache)
- Use sketches for statistical guideposts (Foresight)
Challenge: APQ for Visual Analytics II

**APQ and perception**
- Not well understood
- Need theory and user studies
- Need collaboration with HCI community

A bad visualization [Few 2007]

A bad interface [Fisher et al. 2012]
Challenge: APQ for Visual Analytics III

**Visualizing uncertainty**
- Needed to engender trust, ensure proper inferences
- Don’t need precision < screen resolution [Jugel, et al. 2014]

---

Finite-population confidence bands

Resampling [Kwon et al. 2017]

CLOUDS [Hellerstein et al. 1999]
Challenge: APQ for Visual Analytics IV

Visualizing sample quality

- Helpful for building trust [Fisher et al. 2012]

- Interactive steering of sampling process [Kwon et al. 2017]
Other Challenges

Combining synopses
- Ex: count-min sketch $\rightarrow l_2$-sample $\rightarrow$ estimate of $F_2$

End-to-end incorporation of risk
- Data analysis for decision making under uncertainty
- Choose accuracy of approximation to control risk

Handling Multiple types of uncertainty
- Ex: AQP in probabilistic databases
- Ex: Gaussian random field interpolation

$I_2$-sample: return $(I, R)$, where

$$\Pr(I = i) = (1 \pm \varepsilon) \frac{f_i^2}{F_2} \text{ and } R = (1 \pm \varepsilon) f_i$$
A Random Sample of References

APQ SYSTEMS
- BlinkDB: Queries with bounded errors and bounded response times on very large data. Agarwal et al., Eurosys 2015.
- Quickr: Lazily approximating complex ad hoc queries in bigData clusters. Kandula et al., SIGMOD 2016.

SAMPLING
- One sketch for all: Theory and application of conditional random sampling. Li et al., NIPS 2008.
- Temporally-biased sampling for online model management. Hentschel et al., EDBT 2018.

MISCELLANEOUS
- Neighbor-sensitive hashing. Park et al., VLDB 2015.
References, Continued

APQ FOR VISUAL ANALYTICS

- Visualization-aware sampling for very large databases. Park et al., *ICDE* 2016.

APQ SYNOPSISES: SURVEYS AND COMPARISONS


LEARNING AND BAYESIAN SYNOPSISES

- Database learning: Toward a database that becomes smarter every time. Park et al., *SIGMOD* 2017.