CrowdDB: Answering Queries with Crowdsourcing

Michael Franklin, Tim Kraska, Andrew Wang, Reynold Xin (UC Berkeley)

Donald Kossmann, Sukriti Ramesh (ETH Zurich)

Sam Madden (MIT)
Who am I?

- Reynold Xin 辛湜 (shi2)
- UC Berkeley PhD student, 2nd year
- Google Research / Ads Infrastructure Engineering
- IBM Distributed Database Engineering
- Altera FPGA Software Engineering
- University of Toronto, Engineering Science
Agenda

• Big Data and the AMPLab (Algorithms, Machines, People)

• Algorithms: Error bars

• Machines: Mesos (A Data-Center OS)

• People: CrowdDB (Answering Queries with Crowdsourcing)

• Questions & Answers
The future is data-based
Big Data is massive

- Facebook: 130 TB/day user logs, 400 TB/day pictures.
- Google: > 25 PB/day processed data.
- Gene sequencing: 100M kilobases per day per machine.
- Total data created in 2010: 1 ZettaByte (1,000,000 PB)
- 60% increase every year!
“Big Data”: working definition

When the normal application of current technology doesn’t enable users to obtain timely answers of sufficient quality to their data-driven questions.
Algorithms, Machines, and People
AMP Lab: Making sense at scale by integrating Algorithms, Machines, and People
AMPLab: What is it?

A five-year research collaboration to develop a new generation of data analysis methods, tools and infrastructure for making sense at scale.

(Started Feb 2011)
### The AMP Team: Principle Investigators

<table>
<thead>
<tr>
<th>Name</th>
<th>Research Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex Bayen</td>
<td>sensing platforms (civil eng)</td>
</tr>
<tr>
<td>Armando Fox</td>
<td>systems</td>
</tr>
<tr>
<td>Michael Franklin*</td>
<td>databases</td>
</tr>
<tr>
<td>Michael Jordan*</td>
<td>machine learning</td>
</tr>
<tr>
<td>Anthony Joseph</td>
<td>security &amp; privacy</td>
</tr>
<tr>
<td>Randy Katz</td>
<td>systems</td>
</tr>
<tr>
<td>David Patterson</td>
<td>systems</td>
</tr>
<tr>
<td>Ion Stoica*</td>
<td>systems</td>
</tr>
<tr>
<td>Scott Shenker</td>
<td>networking</td>
</tr>
</tbody>
</table>

* co-directors
AMPLab sponsors
Crowdsourcing is the act of outsourcing tasks, traditionally performed by an employee or contractor, to an undefined, large group of people or community, through an open call.

- Wikipedia
You're invited to Wikipedia Loves Libraries, a program of events at libraries and archives across North America around October 2011.

Crowdsourcing

From Wikipedia, the free encyclopedia

Crowdsourcing is the act of sourcing tasks traditionally performed by specific individuals to an undefined large group of people or community (crowd) through an open call. (FZS)

Jeff Howe established that the concept of crowdsourcing depends essentially on the fact that because it is an open call to an undefined group of people, it gathers those who are most fit to perform tasks, solve complex problems and contribute with the most relevant and fresh ideas.

For example, the public may be invited to develop a new technology, carry out a design task (also known as community-based design[^1] or "design by democracy" and distributed participatory design), refine or carry out the steps of an algorithm (see human-based computation), or help capture
steamboat train, from New this morning ran off the track New-London. Four cars plunged

reCAPTCHA

Implicit Crowdsourcing
Farm Ville
Implicit Crowdsourcing
人肉搜索引擎

人肉搜索引擎，是指利用人工参与来提纯搜索引擎提供信息的一种机制，实际上是通过其他人来搜索自己搜不到的东西，与知识搜索的概念差不多，只是更强调搜索过程的互动而已。搜索引擎也有可能对一些问题不能进行解答，当用户提的问题在搜索引擎中找不到答案时，就会试图通过其他人与人的沟通交流寻求答案。
Amazon Mechanical Turk (Virtualized Humans on Demand)

Mechanical Turk is a marketplace for work.
We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it’s convenient.

79,067 HITs available. View them now.

Make Money by working on HITs
HITs - Human Intelligence Tasks - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:
- Can work from home
- Choose your own work hours
- Get paid for doing good work

Get Results from Mechanical Turk Workers
Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. Register Now

As a Mechanical Turk Requester you:
- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

Find an interesting task
Work
Earn money

Fund your account
Load your tasks
Get results

Find HITs Now
Get Started

or learn more about being a Worker

FAQ | Contact Us | Careers at Amazon | Developers | Press | Policies | Blog
©2005-2011 Amazon.com, Inc. or its Affiliates

An amazon.com company
Amazon Mechanical Turk

- Amazon Mechanical Turk API

  - Requestors place Human Intelligence Tasks (HITs) via “createHit()” API

  - Parameters include: #of replicas, expiration, User Interface,…

- Requestors approve jobs and payment

  - “getAssignments()”, “approveAssignments()”

- Workers (a.k.a. “turkers”) choose jobs, do them, get paid
## Workers’ View

### Table of HITs

<table>
<thead>
<tr>
<th>HIT Details</th>
<th>Requester</th>
<th>HIT Expiration Date</th>
<th>Time Allotted</th>
<th>Reward</th>
<th>HITs Available</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>huge test</strong></td>
<td>Mr. Doe</td>
<td>Aug 27, 2011 (1 day 5 hours)</td>
<td>60 seconds</td>
<td>$0.01</td>
<td>8424</td>
</tr>
<tr>
<td><strong>Validate Brand/Product Information from Product Picture</strong></td>
<td>Redwood Technologies</td>
<td>Aug 27, 2011 (11 hours 59 minutes)</td>
<td>15 minutes</td>
<td>$0.01</td>
<td>6</td>
</tr>
<tr>
<td><strong>Copy Brand/Product Information from Product Picture</strong></td>
<td>Redwood Technologies</td>
<td>Aug 27, 2011 (11 hours 58 minutes)</td>
<td>15 minutes</td>
<td>$0.03</td>
<td>8</td>
</tr>
<tr>
<td><strong>huge job test for refactored balancer</strong></td>
<td>Mr. Doe</td>
<td>Aug 27, 2011 (1 day 5 hours)</td>
<td>60 seconds</td>
<td>$0.01</td>
<td>12390</td>
</tr>
</tbody>
</table>
Workers’ View

Are these two pictures of the same person?

Yes ☐ No ☐

Please ACCEPT the hit before submitting.
Requesters’ View
Motivation

SQL> SELECT * FROM pictures WHERE topic=`Golden Gate Bridge'
    ORDER BY relevance

RDBMS: ?????????????????????????????
Motivation

<table>
<thead>
<tr>
<th>company name</th>
<th>address</th>
<th>market cap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>Mountain View, CA</td>
<td>150</td>
</tr>
<tr>
<td>Microsoft</td>
<td>Redmond, WA</td>
<td>200</td>
</tr>
<tr>
<td>International Business Machines</td>
<td>Armonk, NY</td>
<td>150</td>
</tr>
</tbody>
</table>

SELECT Market_Cap FROM Companies
WHERE Company_Name = “IBM”

Number of rows: 0

Problem: Entity Resolution
Motivation

<table>
<thead>
<tr>
<th>company name</th>
<th>address</th>
<th>market cap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>Mountain View, CA</td>
<td>150</td>
</tr>
<tr>
<td>Microsoft</td>
<td>Redmond, WA</td>
<td>200</td>
</tr>
<tr>
<td>International Business Machines</td>
<td>Armonk, NY</td>
<td>150</td>
</tr>
</tbody>
</table>

SELECT Market_Cap FROM Companies
WHERE Company_Name = “Apple”

Number of rows: 0

Problem: Closed World Assumption
Hybrid Human-Machine DBMS

- Hard database problems
  - Missing data
  - Fuzzy comparisons

- CrowdDB approach
  - Let machines do what machines do best
  - Let people do what people do best
A radical new idea?

“The hope is that, in not too many years, human brains and computing machines will be coupled together very tightly, and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today.”
CrowdDB and CrowdSQL

create table company (  
    name string,  
    address CROWD string  
);

create CROWD table department (  
    name string primary key,  
    reception_phone_number string  
);

select address from company  
where name = "Big Blue"

select p from pictures  
where subject = "Golden Gate Bridge"  
order by  
CROWDORDER(p, "better visualize ...");
User Interface Generation

• A clear UI is key to response time and answer quality.

• Can leverage the SQL Schema to auto-generate UI (e.g., Oracle Forms, etc.)
Query Execution

(a) PeopleSQL query
(b) Logical plan before optimization
(c) Logical plan after optimization
(d) Physical plan
Query Execution

(b) Logical plan before optimization
(c) Logical plan after optimization
(d) Physical plan

Rule based optimizer
- Performance Insightful Query Language (PIQL) techniques to deal with open-world assumption
- Simple set of rules to pick the best plan
- Simple heuristics to set the crowd parameters (e.g., replication factor, price per HIT, etc.)
Query Optimization and Execution

- Creates user interface templates
- Select physical operators
- New Query Operators:
  - Crowd Operators: MTProbe, MTJoin, MTFuncti
  - Other: STOP AFTER (i.e., limit)
CrowdProbe Operator

- Similar to a table-scan with predicate push-downs
- Batches several jobs into one HIT
- Issues as many requests in parallel as possible (based on the cardinality prediction)
- Does simple quality control (quorum votes)
- “Caches” the result inside the corresponding table → queries have side-effects

Crowd Column & Crowd Columns
21 w/o foreign keys

Crowd Column & Crowd Columns
with foreign keys

Denormalization

Please fill out the missing company data!
Name: IBM
Headquarter address: 
Submit

Please fill out the missing professor data
Name: Carey
Department name: CS
E-Mail: 
Submit

Please fill out the missing professor data
Name: Carey
E-Mail: 
Department: 
Department Phone: 
Submit
CrowdJoin and CrowdFunction Operators

**MTJoin**
- Indexed nested-loop join
- Rest (quality control, HIT grouping) similar to MTProbe

**MTJoin**
(Dep)
\[ p.\text{dep} = d.\text{name} \]

**MTProbe**
(Professor)
\[ \text{name} = \text{Carey} \]

**MTFunction**
- implements the CROWDEQUAL and CROWDORDER comparison
- Takes some description and a type (equal, order) parameter
- Quality control again based on majority vote
- Ordering can be further optimized (e.g., Three-way comparisons vs. Two-way comparisons)

Are the following entities the same?

IBM == Big Blue

Which picture visualizes better "Golden Gate Bridge"?

IBM == Big Blue

Submit

Friday, February 24, 12
User Interface vs Quality

MTJoin (Professor)  
p.name = "carey"

MTJoin (Dep)  
p.dep = d.name

MTProbe (Dep)  
Please fill out the missing department data
Department Name
Phone
Submit

MTProbe (Professor)  
name=Carey

MTProbe (Professor, Dep)  
name=Carey

Department first
Professor first
10% error rate
De-normalized Probe
(80% error rate)
Does it work?
Entity Resolution

**Schema:**
```
CREATE TABLE company ( 
  name STRING,  
  headquarter_address CROWD STRING  
);
```

**Query:**
```
SELECT name  
FROM company  
WHERE name ~ [a non-uniform  
  name of the company]
```

### Which entities are the same as **Big BLUE?**
- Google
- HP
- IBM
- Facebook
- NetApp
- CrowdFlower
- Yahoo
- Microsoft
- Salesforce
- SAP
- None of the above

---

<table>
<thead>
<tr>
<th>Non Uniform Name</th>
<th>Query Result</th>
<th>Votes</th>
<th>Error Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayerische Motoren Werke</td>
<td>BMW</td>
<td>3</td>
<td>TATA Group, Gazprom, Boeing, Toyota</td>
</tr>
<tr>
<td>International Business Machines</td>
<td>IBM</td>
<td>2</td>
<td>Samsung, HP</td>
</tr>
<tr>
<td>Company of Gillette</td>
<td>P&amp;G</td>
<td>2</td>
<td>Aviva, AIG, France Telecom</td>
</tr>
<tr>
<td>Big Blue</td>
<td>IBM</td>
<td>2</td>
<td>Microsoft</td>
</tr>
</tbody>
</table>

Data-Size: 100 company names  
Batching: 10 comparisons per HIT  
Replication: 3 Assignments per HIT  
Price: 1 cent per HIT
Query Optimization

• SQL is declarative. You specify what you want, not how you want it.

• SELECT department.name, professor.phone_number FROM department, professor WHERE professor.name="Mike Carey" AND department.id=professor.department_id

• Data independence: Data outlive applications.
The Dream: A Cost-Based Optimizer

- SQL provides logical and physical data independence.

- Crowdsourcing platforms have so many parameters to tweak. Impossible for programmers to get them all right.

- Parameters (market dynamics) change over time.

- Can we create a cost model (monetary, time, and quality) for crowd operators and plug them into the database optimizer?
Can we build a crowd optimizer?

Select *
From Restaurant
Where city = ...

Friday, February 24, 12
An Example Benchmark

![Graph showing % of HITs completed vs Price]

- Time (mins)
- Percentage of HITs completed

Curves for different prices:
- $0.04
- $0.03
- $0.02
- $0.01
Worker Affinity and Errors

![Graph showing Worker Affinity and Error Rate]

- **X-axis:** Turkers
- **Y-axis:** Number of assignments submitted and Incorrect assignments completed
- **Legend:**
  - Blue line: Total HITs submitted
  - Red line: Incorrect assignments completed
Unexpected Market Behavior

The graph illustrates the relationship between the number of HITs in a HIT group and the percentage of HITs completed. As the number of HITs in a group increases, there is an initial increase in the percentage of HITs completed, reaching a peak around 23.5% at around 25 HITs. Beyond this point, the percentage of HITs completed decreases sharply, indicating a decline in participation or completion rates with larger group sizes. This suggests that larger HIT groups may not be as effective in terms of completing tasks as smaller groups.
The market is smaller than expected

Start of a concurrent experiment

HITs completed (%) vs. Time (mins)

- Y-axis: HITs completed (%)
- X-axis: Time (mins)
Can we build a crowd optimizer?

Select *
From Restaurant
Where city = ...

be very wary of doing any work for this requester...

I advise not clicking on his “information about restaurants” hits.

Hmm... I smell lab rat material.

I would do work for this requester again.

This guy should be shunned.

I would do work for this requester again.
Tim Klas Kraska

HIT Group » I recently did 299 HITs for this requester.... Of the 299 HITs I completed, 11 of them were rejected without any reason being given. Prior to this I only had 14 rejections, a .2% rejection rate. I currently have 8522 submitted HITs, with a .3% rejection rate after the rejections from this requester (25 total rejections). I have attempted to contact the requester and will update if I receive a response. Until then be very wary of doing any work for this requester, as it appears that they are rejecting about 1 in every 27 HITs being submitted.

posted by ...
## Crowd as Specialized CPUs

<table>
<thead>
<tr>
<th>Feature</th>
<th>Cloud</th>
<th>Crowd</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>cost</strong></td>
<td>$0.02 - $2.10 / hr</td>
<td>$4.8 / hr</td>
</tr>
<tr>
<td><strong>pay model</strong></td>
<td>pay as you go</td>
<td>pay as you go</td>
</tr>
<tr>
<td><strong>investment</strong></td>
<td>none</td>
<td>training</td>
</tr>
<tr>
<td><strong>response time</strong></td>
<td>varied: msecs</td>
<td>varied: mins, hours, days</td>
</tr>
<tr>
<td><strong>capacities</strong></td>
<td>good at number crunching</td>
<td>good at “AI”</td>
</tr>
<tr>
<td><strong>programming</strong></td>
<td>formal</td>
<td>natural language &amp; UI</td>
</tr>
<tr>
<td><strong>availability</strong></td>
<td>virtually unlimited</td>
<td>???</td>
</tr>
<tr>
<td><strong>affinity</strong></td>
<td>virtualized</td>
<td>personal</td>
</tr>
<tr>
<td><strong>reliability</strong></td>
<td>faulty but not malicious</td>
<td>can be malicious</td>
</tr>
<tr>
<td><strong>legal</strong></td>
<td>privacy</td>
<td>privacy, taxes, benefits...</td>
</tr>
</tbody>
</table>
Crowdsourcing Research Challenges

• High quality results from low quality sources

• Batch processing (prefetching) vs real-time

• Spammers and bots

• Confidentiality

• Privacy and other moral & legal issues
Summary - CrowdDB

• A first attempt towards the P in AMP.

• CrowdDB is a hybrid Crowd/Cloud computing
  
  • Small set of SQL extensions allow to express how the crowd should be used
  
  • Special crowd operators encapsulate the input of the crowd
  
  • Shows that people can help answer DB-hard Queries
  
  • And, it raises lots of interesting and important research issues.
Thank You!

Questions?

rxin@cs.berkeley.edu

http://www.cs.berkeley.edu/~rxin/