Relational Data Markets in the Cloud: Challenges and Opportunities

Magdalena Balazinska

Joint work with Bill Howe, Dan Suciu, and the database group

University of Washington

Project supported in part by NSF and Microsoft

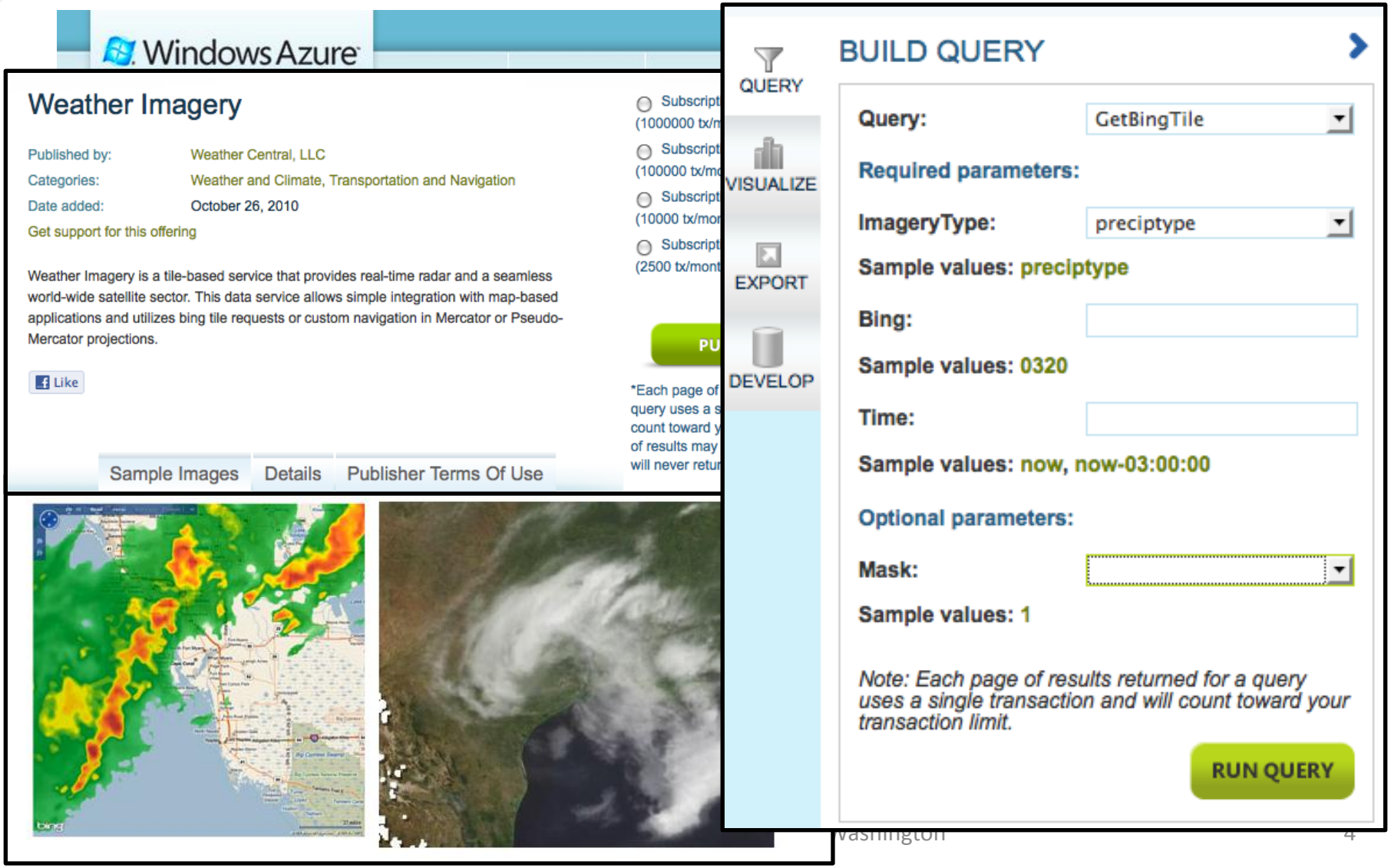
Cloud Data Markets

- Data is valuable!
  - Real-time stock prices + trade data: $35,000/year (https://www.xignite.com)
  - Parcel information: $60,000/year (https://datamarket.azure.com)
- Data can help businesses and application developers
- Cloud can facilitate buying and selling data

Data Markets

- Single, logically centralized point for buying/selling data, often facilitated by cloud computing
  - datamarket.azure.com
  - www.infochimps.com
  - www.customlists.net/home
  - www.aggdata.com
  - …

Azure DataMarket



Technical Challenges

- **How to price data?**

- What tools does a data market need?

Current Data Pricing Models

- Subscription based (Azure DataMarket):
  - Fixed number of *transactions* per month
  - Each transaction = 100 records

- Fixed price for the entire dataset (Customlist)
  - Data owners define subsets and sell at lower price

Canadian Business Database

www.customlists.net/canadian-business-database

CustomLists.net
YOUR #1 DATA SOURCE!

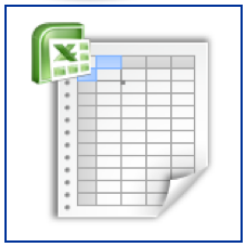
Call us on 1800 495 9313

Click here for »
Live Help

VeriSign Trusted
VERIFY

Home | Available Databases | Guarantee | FAQ | About Us | Contact Us

Canadian Business Database



Our Price: ~~USD$699.00~~ **Special Price: USD$399.00**

Buy Now

CustomLists.net is proud to bring you the Canadian Business Database. This is the most comprehensive database of its kind available and is also one of the most up-to-date with data continually updated on a daily basis.

The Most Comprehensive Database of Canadian Businesses:

- 2.3 million Listings
- 410,638 Email Addresses
- 1.9 million Contact Names
- 2.2 million Phone Numbers
- 834,525 Fax Numbers
- 2.3 million Addresses

All Data Verified and Cross-Checked via:

- Yellow Pages Directory
- White Pages Directory
- CRA Records
- Company Registration Records

Listings Include:

- Business Category

Download a FREE
Sample Database

Simply enter your email and select a database to receive a free sample via email.
(Multiple selections are allowed).

Your Email Address

Download Now

We respect your privacy: we never send spam email and your email address will never be shared with anyone else.

100% SATISFACTION GUARANTEED

100% Satisfaction Guarantee

Client Results:

"Fast, prompt and friendly service. We will continue to use CustomLists.net for our clients marketing lists."
Michael Sanders

Cheaper by province

Alberta Business Database

www.customlists.net/alberta-business-database

CustomLists.net
YOUR #1 DATA SOURCE!

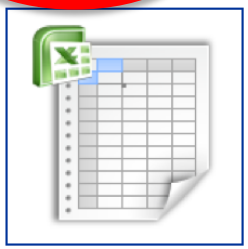
Call us on 1800 495 9313

Click here for »
Live Help

VeriSign Trusted
VERIFY

Home | Available Databases | Guarantee | FAQ | About Us | Contact Us

Alberta Business Database



Our Price: ~~USD$299.00~~

Special Price: USD$199.00

Buy Now

CustomLists.net is proud to bring you the Alberta Business Database. This is the most comprehensive database of its kind available and is also one of the most up-to-date with data continually updated on a daily basis.

The Most Comprehensive Database of Canadian Businesses:

- 290,812 Listings
- 82,445 Email Addresses
- 246,820 Contact Names
- 277,327 Phone Numbers
- 151,745 Fax Numbers
- 290,812 Addresses

All Data Verified and Cross-Checked via:

- Yellow Pages Directory
- White Pages Directory
- CRA Records
- Company Registration Records

Listings Include:

- Business Category

Download a FREE
Sample Database

Simply enter your email and select a database to receive a free sample via email.
(Multiple selections are allowed).

Your Email Address

Download Now

We respect your privacy: we never send spam email and your email address will never be shared with anyone else.

100% SATISFACTION GUARANTEED

100% Satisfaction Guarantee

Client Results:

"I recently ran an email campaign using the American Business Email Database purchased from CustomLists.net The cost of the campaign was paid for within

Data Pricing Issues

- Lump sum or subscription pricing is **inflexible**
  - For lump sum, can only buy pre-defined views
  - For subscription, can only ask pre-defined queries

- Not clear how to price **updates**

Data Pricing Issues (continued)

- Today's data pricing can have **bad properties**
- EX: Weather Imagery on Azure DataMarket
  - 1,000,000 transactions -> $2,400
  - 100,000 -> $600
  - 10,000 -> $120
  - 2,500 -> $0
- **Arbitrage opportunity**:
  - Emulate many users
  - Get as much data as you want for free!

Data Pricing Issues (continued)

- Not clear how to price the output of queries that **correlate multiple datasets**
- And how to support queries **across data owners**?
- **No principled method for pricing data**

Alternative Pricing Models

- We are studying different pricing models
  - Pricing using the PRICE-semiring
  - Pricing using provenance expressions
  - View-based pricing

- Goal: address pricing challenges through
  - Fine-grained data pricing
  - Automatic price derivations for complex queries
  - Price extrapolation for new queries

Pricing Using PRICE-Semiring

- Approach
  - Assign a price to individual base tuples
  - Automatically compute the price of query result

R

A	B	C	
a	b	e	p
d	b	g	r
a	c	e	s

S

B	D	
b	x	q
c	x	t
b	y	u

Q = `SELECT DISTINCT A,D FROM R,S WHERE R.B = S.B AND S.D=x`

A	D	
a	x	$min(p + q, s + t)$
d	x	$r + q$

a pricing function on tuples:

$p = \$0.1$ $q = \$0.02$
$r = \$0.01$ $t = \$0.03$
$s = \$0.5$ $u = \$0.04$

a pricing calculation:

$min(p + q, s + t) = \$0.12$
$r + q = \$0.03$

$price(Q) = \$0.15$

Pricing Using PRICE-Semiring

- Benefits
  - Assign different prices to different tuples
  - Allow users to ask arbitrary queries
  - Can compute prices across datasets and owners
  - Avoid bad properties such as arbitrage

Pricing Using Provenance Expressions

- Approach: Same as above BUT
  - Derive provenance information for each result tuple
  - Price is a function of these provenance tuples

R

A	B	C	
a	b	e	*p*
d	b	g	*r*
a	c	e	*s*

S

B	D	
b	x	*q*
c	x	*t*
b	y	*u*

Q =
```
SELECT DISTINCT A,D FROM R,S
WHERE R.B = S.B AND S.D=x
```

A	D	
a	x	*Provenance: p, q, s, and t*
d	x	*Provenance: r, q*

a pricing function on tuples:

p = $0.1
r = $0.01
s = $0.5
q = $0.02
t = $0.03
u = $0.04

a pricing calculation (applying a 25% discount):

price(Q) = f(price(p), price(q), price(r), price(s), price(t))
= 0.75 (0.1 + 0.02 + 0.01 + 0.5 +0.03)
= $0.50

Pricing Using Provenance Expressions

- Benefits
  - Assign different prices to different tuples
  - Allow users to ask arbitrary queries
  - Can compute prices across datasets and owners
  - More powerful pricing functions become possible
    - E.g., submodular pricing
- But properties with complex pricing needs studying
- Naïve implementation could be inefficient

View-Based Pricing

- Approach
  - User specifies a set of queries Q_1, … Q_n
  - And their prices: price(Q_1), …, price(Q_n)
    - D = all businesses in North America
    - V_1 (businesses in Canada) = $600
    - V_2 (businesses in Alberta) = $300
    - V_3 (all Shell businesses) = $50
    - Etc.
  - System automatically computes other query prices
    - Q = find all businesses with > 200 employees. What's the price of this data ?

View-Based Pricing (continued)

- This is a constrained optimization problem
  - Each query price is a constraint
  - Can add other constraints: e.g., total price of DB

- Two methods to derive prices of new queries
  - Reverse-eng. price of base tuples s.t. constraints
    - Assume a function that converts base tuple prices into query prices
    - Compute base tuple prices in a way that maximizes entropy, user utility, or other function s.t. constraints
  - Compute new query prices directly

View-Based Pricing

- Benefits
  - Assign different prices to different tuples
  - Allow users to ask arbitrary queries
  - Can compute prices across datasets and owners
  - Most powerful pricing schemes become possible
  - Enables the creation of different versions for different users
  - Need to build pricing functions enforcing desired properties

Technical Challenges

- How to price data?

- **What tools does a data market need?**

Data Market Tools

- Efficient query-price computer
  - Fine-grained prices should not add much overhead
- Update provider
  - Given an earlier user-query with a price
  - Compute delta query output after DB update
  - Compute price of update

Data Market Tools (continued)

- Pricing Advisor
  - Checks properties of a pricing scheme
  - Helps tune prices based on data provider goals
  - Computes prices of new views
  - Explains income or bill
  - Etc.

- Price-aware query optimizer
  - Answer query over multiple datasets as cheaply as possible

Conclusion

- Data helps drive businesses and applications
- Data markets are emerging, facilitated by cloud
- But need the right tools to maximize success
  - Theory of data pricing
  - Systems for computing prices, checking properties, etc.
- Stay tuned for more:
  http://data-pricing.cs.washington.edu