MCDB and SimSQL: Scalable Stochastic Analysis within the Database

Peter Haas (IBM) and Chris Jermaine (Rice University)

People who have worked on MCDB/SimSQL: Subi Arumugam, Zhuhua Cai, Jacob Gao, Shangyu Luo, Ravi Jampani, Luis Perez, Foula Vagena, Mingxi Wu, Fei Xu

MCDB/SimSQL: What Is It?

- Complete analytic SQL DB system
- Runs on top of Hadoop
- Scales easily to 100-machine clusters
- Handles terabytes of data
- Special support for ***stochastic analytics***

What Does "Stochastic Analytics" Mean?

- Data analytics based on stochastic simulation
- Model missing/uncertain data via prob. dists
- System uses prob. dists to simulate DB instances
- Can then query those instances via SQL
- Complicated dists supported, even Markov chains

Special Features of MCDB/SimSQL

- Unique randomized table functions ("VG functions")
- Multiple concurrent simulations for the price of one
- Special support for recursion (MCMC and Bayesian ML)
- Support for incremental execution of very large plans
- Coming soon: first class vectors and matrices

Simple example: Writing SQL queries that take into account errors on sensor readings.

LOCATIONS

SensID	GrpID	Lat	Long
1	1	21.410	42.317
2	1	21.419	42.336
3	1	21.397	42.346
4	2	21.678	42.382
5	2	21.746	42.637
6	3	21.312	42.551
7	3	21.287	42.528
8	3	21.300	42.574
9	3	21.307	42.559
10	4	21.555	42.163
11	4	21.572	42.201

READINGS

SensID	Time	Temp
1	10:13:03	82.77
2	10:13:03	83.62
3	10:13:03	79.22
3	10:13:22	79.25
4	10:13:03	71.82
5	10:13:03	74.27
5	10:13:22	74.27
6	10:13:03	77.37
6	10:13:22	77.38
6	10:13:45	77.43
7	10:13:03	79.00

COVARS

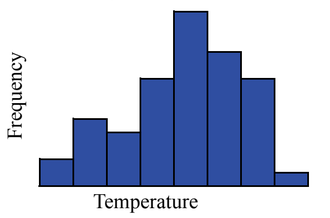
SensA	SensB	Covar
1	1	1.21
2	1	0.45
3	1	0.72
2	2	0.75
2	3	0.63
3	3	1.10
4	4	0.55
5	4	0.37
5	5	0.95
6	6	1.00
7	6	0.04

SENSORS

ID	Lat	Long	Temp	Time
1	21.410	42.317	82.93	10:13:03
2	21.419	42.336	82.92	10:13:03
3	21.397	42.346	80.22	10:13:03
3	21.397	42.346	81.35	10:13:22
4	21.678	42.382	72.52	10:13:03
5	21.746	42.637	75.29	10:13:03
5	21.746	42.637	75.29	10:13:22
6	21.312	42.551	76.47	10:13:03
6	21.312	42.551	76.49	10:13:22
6	21.312	42.551	76.47	10:13:45
7	21.287	42.528	75.55	10:13:03

This column is stochastic!

Empirical Distribution of Query Results



Stored Data

"VG Function" to produce stochastic data

```
CREATE TABLE SENSORS(ID, LAT, LONG, TEMP, TIME) AS
FOR EACH g IN (SELECT DISTINCT GrpID, Time FROM Readings, Locations)
    WITH TEMP AS MDNormal(
        (SELECT r.SensID, r.Temp
         FROM Readings r, Locations l
         WHERE l.GrpID = g.GrpID AND l.SensID = r.SensID AND
               r.Time = g.Time),
        (SELECT c.SensA, c.SensB, c.Covar
         FROM Covars c, Locations l
         WHERE c.SensA = l.SensID AND l.GrpID = g.GrpID))
    SELECT l.SensID, l.Lat, l.Long, t.value, g.time
    FROM Locations l, Temp t
    WHERE l.SensID = t.ID
```

prepares mean vector

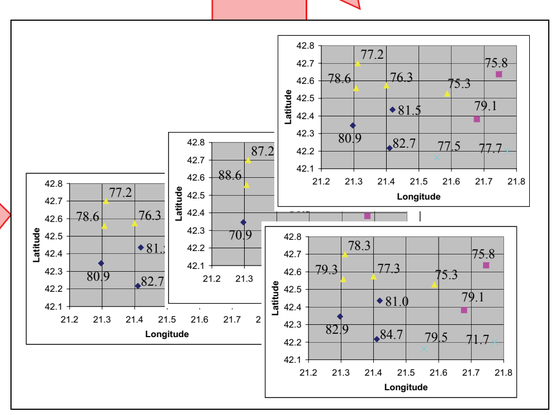
prepares covariance matrix

Parameterization

MDNormal Variable Generation (VG) function:

$$\text{TEMP} \sim \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Simulation



Simulated DB

Simulated DB

Simulated DB

Simulated DB

Query

"For each time tick at which the sensor measurements were recorded, what was the average temperature in the box Lat = [21.5, 21.7], Long = [42.5, 42,7]?"

```
SELECT AVG (Temp), Time
FROM Sensors
WHERE Lat > 21.5 AND Lat < 21.7 AND
      Long > 42.5 AND Long < 42.7
GROUP BY Time
```

Scalable Bayesian Machine Learning.
Stochastic Analytics.
SQL.

simSQL

Download SimSQL today: http://cmj4.web.rice.edu/SimSQL/SimSQL.html

The design and development of MCDB/SimSQL has been supported by the NSF and the DoE