MCDB: The Monte Carlo Database System

Chris Jermaine
Rice U.

Joint work with:
Subi Arumugam
Peter Haas
Luis Perez
Foula Vagena
Cai Zhuhua
…and others…

MCDB

- MCDB is a database system being developed at Rice U.
- Extensive support for the declarative subset of SQL
- Full-blown, cost-based query optimizer
- Compiles queries to MapReduce jobs, run on Hadoop
  - — So it scales to very large data sizes
  - — Maybe not well, but it scales!
- But that's not what's interesting about MCDB
  - — It's the native support for stochastic analytics

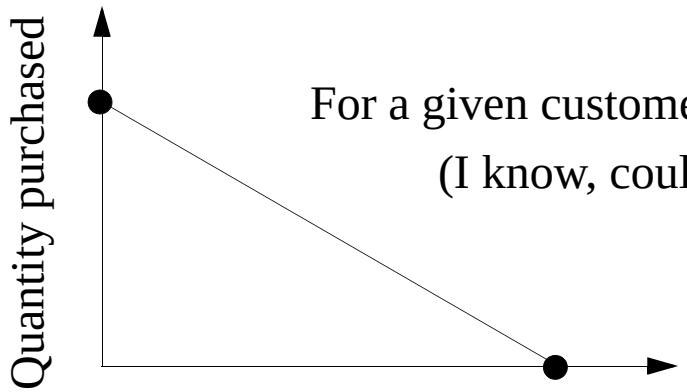
Example

- Say you have archived billions of sales records and want to know:

  "What would my profits have been in '08 if I'd cut all of my margins by 10%?"

- How might we use a data warehouse to guess this answer?
  - — Need to "guess" each customer's demand at new price
  - — Use to build a hypothetical, revised version of sales table
  - — Finally, join this table with others (prices, supply costs, etc.) to compute profits
- Here's how an analyst might use MCDB to do this...

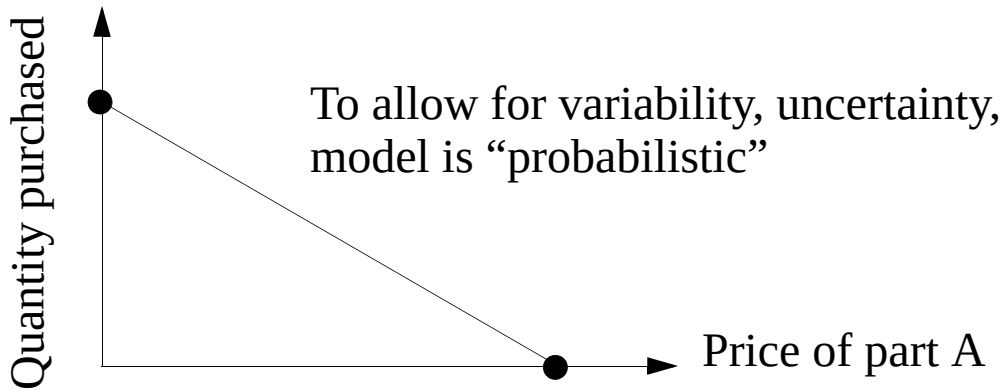
A Stochastic Demand Model



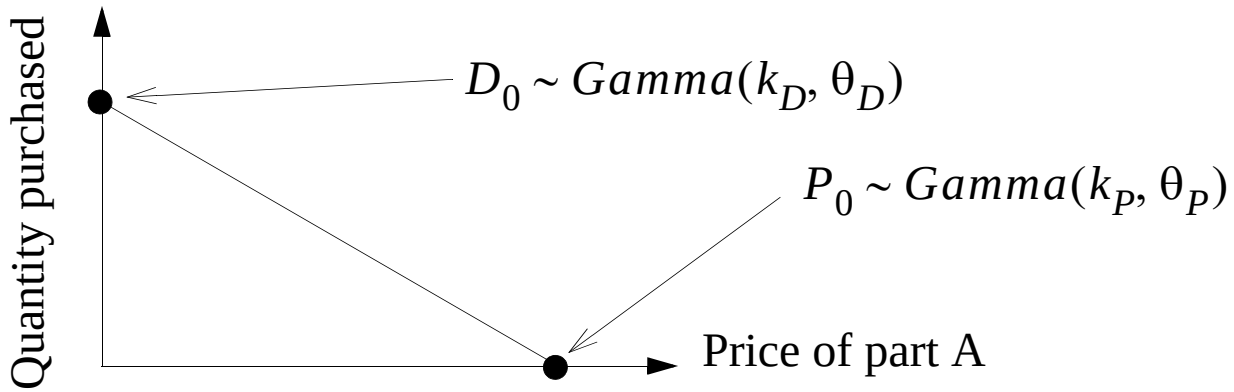
For a given customer, demand is a linear function

(I know, could do better than linear...)

A Stochastic Demand Model

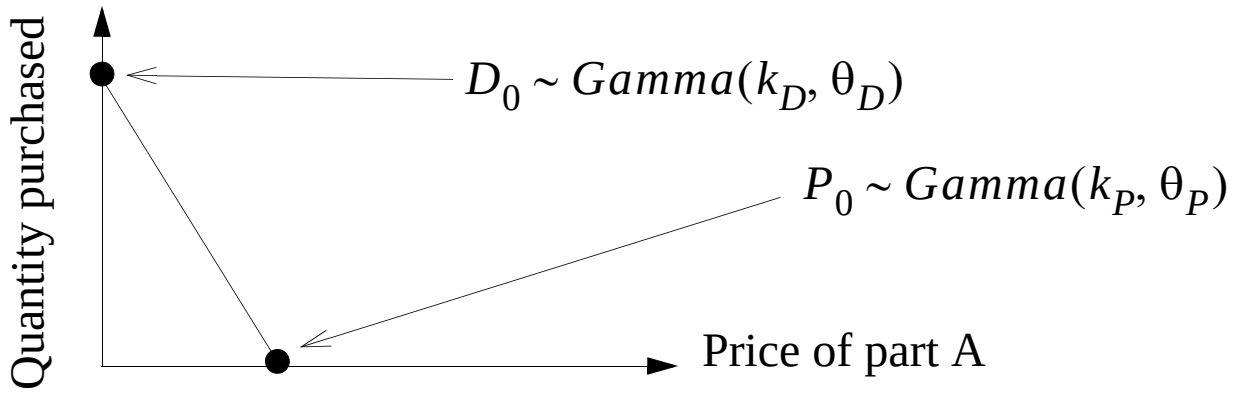


A Stochastic Demand Model



Demand curve is generated via samples from twin Gamma distributions

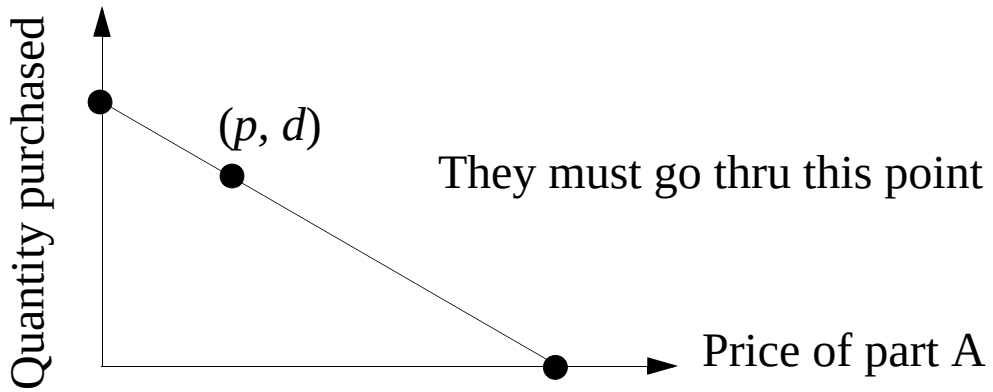
A Stochastic Demand Model



Demand curve is generated via samples from twin Gamma distributions

- This defines what's known as a "prior" over cust demand curves

A Stochastic Demand Model

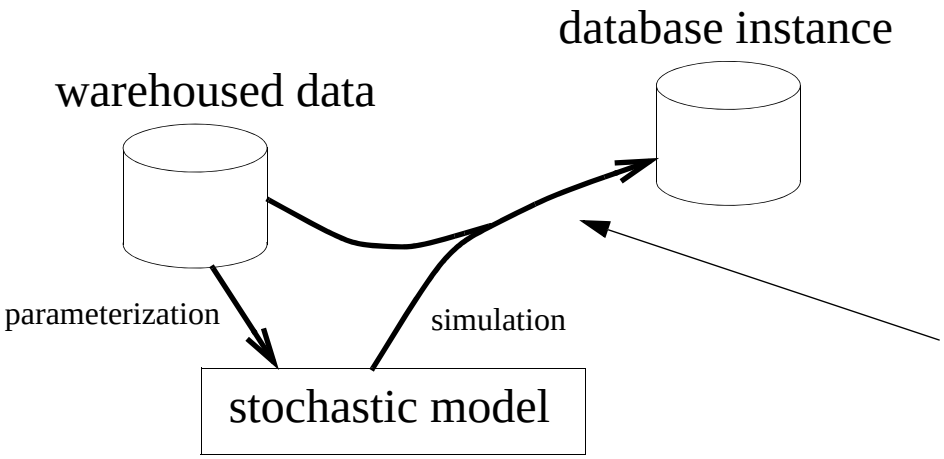


Allowable set of demand curves is a “posterior dist” to a Bayesian

- This defines what’s known as a “prior” over cust demand curves
- Taking into account what the cust actually purchased...
  - — Restricts the set of possible demand curves

This Is Where MCDB Comes In

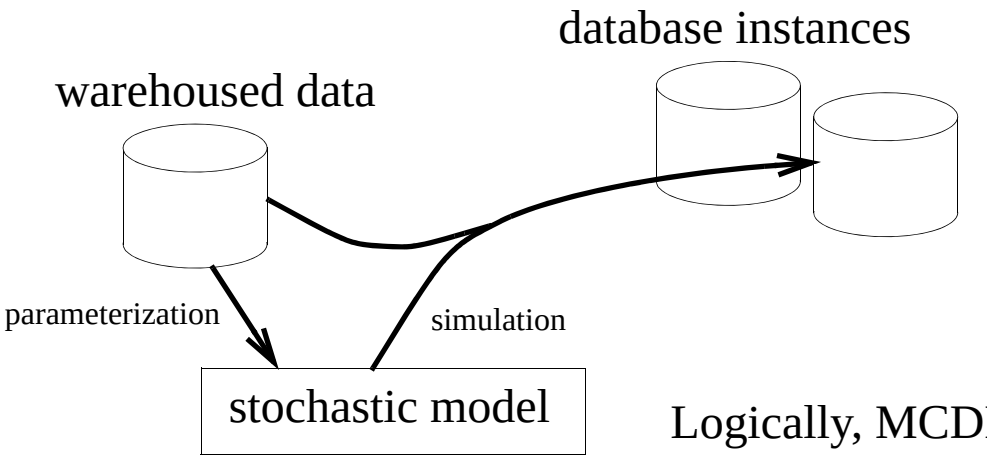
- In MCDB, easy to associate a posterior dist of demand curves...
  - — with every one of the 100M customers in a large database
  - — And then use those curves to generate stochastic DB instances



Done by implementing a “VG Function” which performs the simulation

This Is Where MCDB Comes In

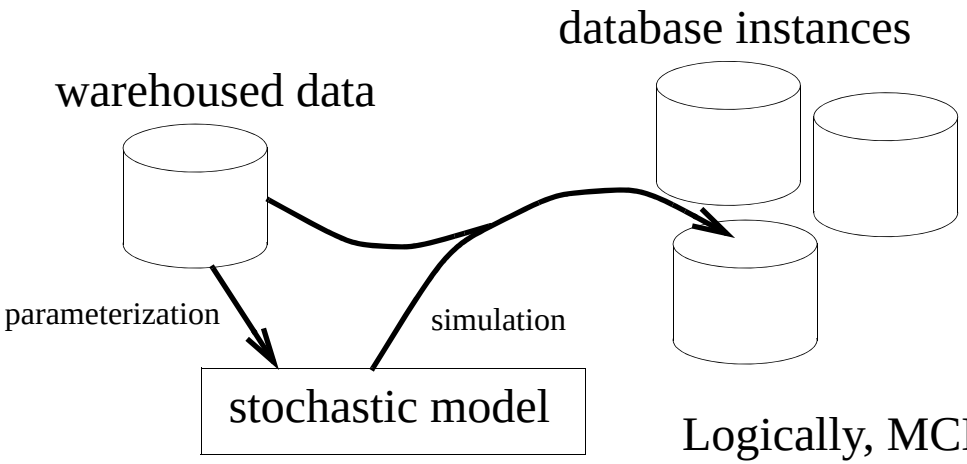
- In MCDB, easy to associate a posterior dist of demand curves...
  - — with every one of the 100M customers in a large database
  - — And then use those curves to generate stochastic DB instances



Logically, MCDB generates many database instances (“possible worlds”)

This Is Where MCDB Comes In

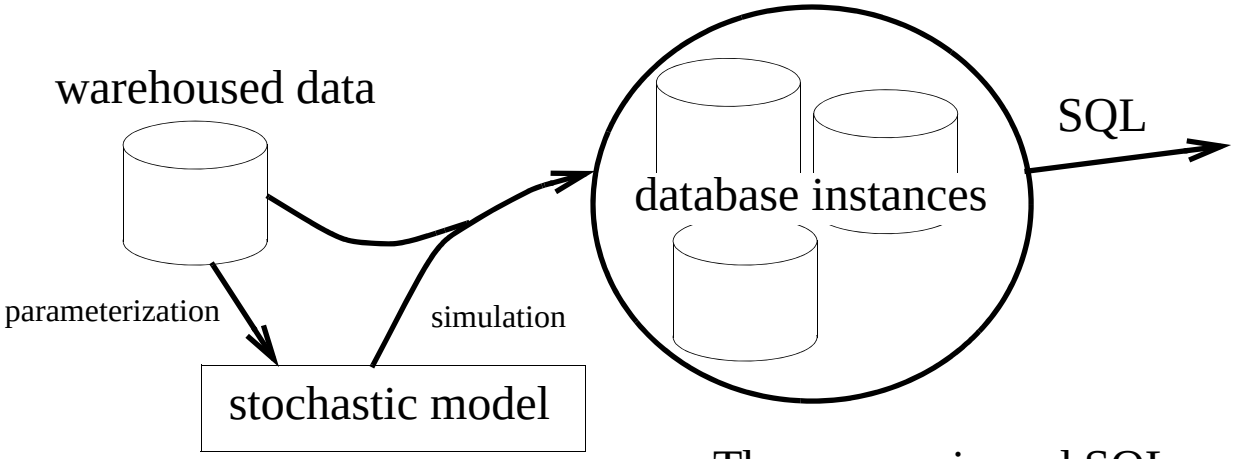
- In MCDB, easy to associate a posterior dist of demand curves...
  - — with every one of the 100M customers in a large database
  - — And then use those curves to generate stochastic DB instances



Logically, MCDB generates many database instances (“possible worlds”)

This Is Where MCDB Comes In

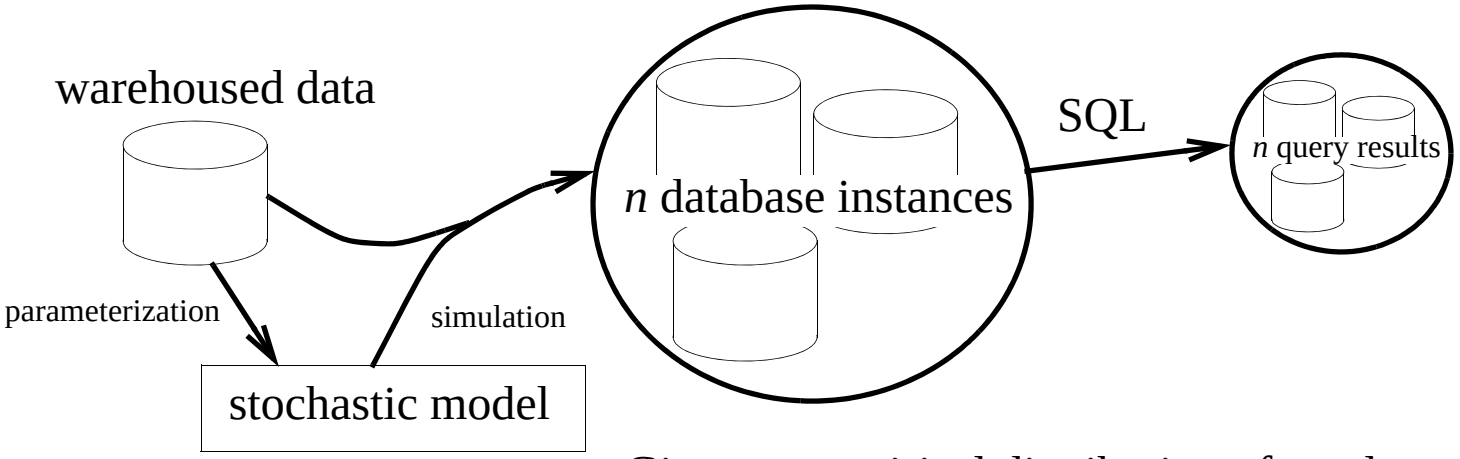
- In MCDB, easy to associate a posterior dist of demand curves...
  - — with every one of the 100M customers in a large database
  - — And then use those curves to generate stochastic DB instances



Then a user-issued SQL query
is simultaneously evaluated over all instances

This Is Where MCDB Comes In

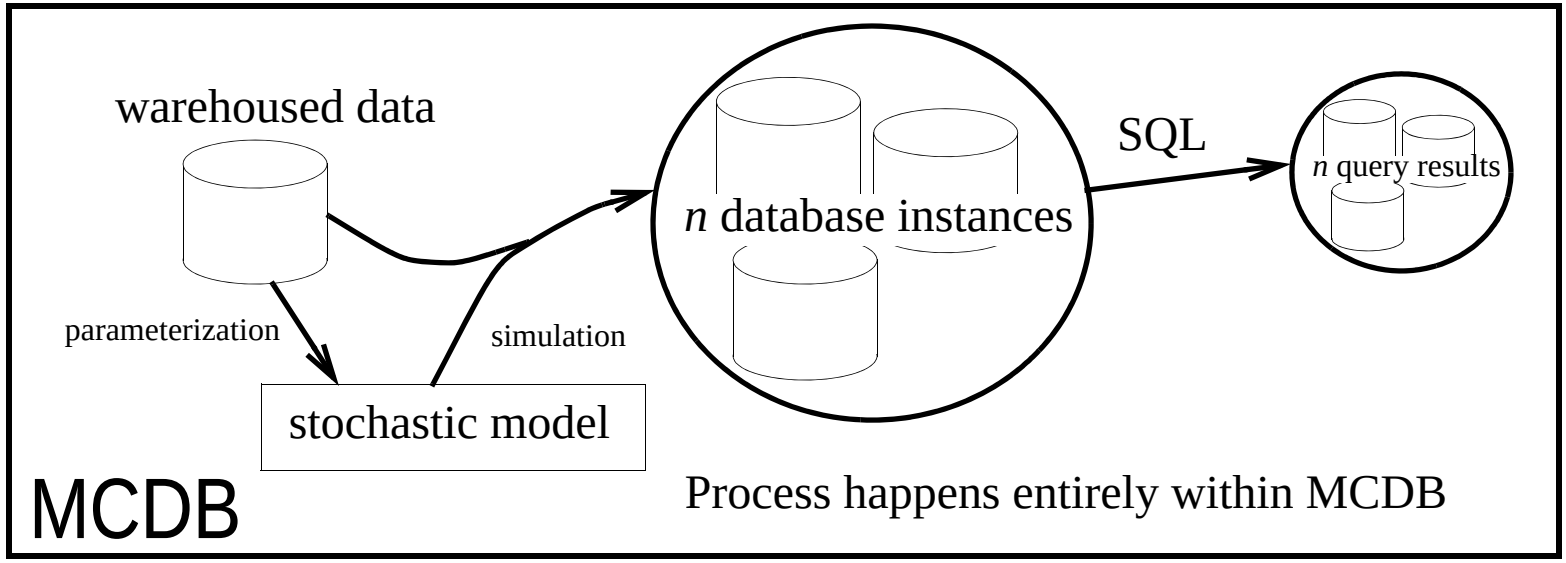
- In MCDB, easy to associate a posterior dist of demand curves...
  - — with every one of the 100M customers in a large database
  - — And then use those curves to generate stochastic DB instances



Gives an empirical distribution of result sets

This Is Where MCDB Comes In

- In MCDB, easy to associate a posterior dist of demand curves...
  - — with every one of the 100M customers in a large database
  - — And then use those curves to generate stochastic DB instances



Download it today!