Scalable Complex Analytics and DBMSs

Michael Stonebraker

DBg Database Group
MIT Computer Science and Artificial Intelligence Lab

Simple Analytics

- SQL operations
  - count, sum, max, min, avg
  - Optional group_by
- Defined on tables
- User interface is Business Intelligence Tools
  - Cognos, Business Objects, …
- Appropriate for traditional business applications

Simple Analytics

- Well served by the data warehouse crowd
- Who are good at this stuff
  - even on petabytes

Complex Analytics

- Machine learning
- Data clustering
- Predictive models
- Recommendation engines
- Regressions
- Estimators

Complex Analytics

- By and large, they are defined on arrays
- As collections of linear algebra operations
- They are not in SQL!
- And often
  - Are defined on large amounts of data
  - And/or in high dimensions

Complex Analytics on Array Data - An Accessible Example

- Consider the closing price on all trading days for the last 20 years for two stocks A and B
- What is the covariance between the two time-series?

$$(1/N) * \text{sum}\ (A_i - \text{mean}(A)) * (B_i - \text{mean}\ (B))$$

Now Make It Interesting …

- Do this for all pairs of 15000 stocks
  - The data is the following 15000 x 4000 matrix

Stock	t_1	t_2	t_3	t_4	t_5	t_6	t_7	….	t_{4000}
S_1									
S_2									
…									
S_{15000}									

Array Answer

- Ignoring the (1/N) and subtracting off the means ….

$$Stock * Stock^T$$

Yabut…..

- Array has 60M cells (easily fits in main memory)
- Multiply complexity is (15000) * (15000) * (2000) floating point operations……
  - .5 Teraflop

- What about hourly data (X8) or tick-level data?
- What about all stocks (X4)?
- What about high, low, volume, …?
- What about bid-ask data?
- What about options?

Gets **big** in a hurry!

System Requirements

- Complex analytics
  - Covariance is just the start
  - Defined on arrays
- Data management
  - Leave out outliers
  - Just on securities with a market cap over $10B
- Scalability to many cores, many nodes and out-of-memory data

These Requirements Arise in Many Domains

- Auto insurance
  - Sensor in your car (driving behavior and location)
  - Reward safe driving (no jackrabbit stops, avoid dangerous intersections)
  - Predict driver risk based on 5000 variables for 1M customers
- Genomics and Healthcare Informatics
  - Look for genes overexpressed in disease populations
  - Create cohort groups for effectiveness studies

These Requirements Arise in Many Domains

- Recommendation engines (people who liked XXX also liked YYY)
  - Clustering customers in a high dimensional space is one popular technique
- Predicting unscheduled down-time in complex machinery (oil refineries, jet engines, helicopters, ….)
  - Predictive modeling in high dimensional spaces

Solution Options

- SAS, R, S, SPSS, …
  - Weak or non-existent data management
- RDBMS
  - Weak or non-existent linear algebra
- 2 Systems
  - Learn 2 systems, and copy the world back and forth
- Hadoop
  - Good only at “embarassingly parallel” tasks
  - Hit the wall the minute you try to scale

Better Answer: An Array DBMS (e.g. Paradigm4/SciDB)

- All-in-one: data management with massively scalable advanced analytics
- Data is updated via time-travel; not overwritten
  - Supports reproducibility for research and compliance
- Supports uncertain data, provenance
- Open source (supported/developed by Paradigm4, Inc.)
- Hardware agnostic

Array Query Language (AQL)

- Array data management
  - e.g. filter, aggregate, join, etc.
- Statistical & linear algebra operations
  - multiply, QR factorization, etc.
  - parallel, disk-oriented
- User-defined operators (Postgres-style)

Array Query Language (AQL)

```
SELECT Geo-Mean ( T.B )
FROM Test_Array T
WHERE
    T.I BETWEEN :C1 AND :C2
AND T.J BETWEEN :C3 AND :C4
AND T.A = 10
GROUP BY T.I;
```

User-defined aggregate on an attribute B in array T

Subsample

Filter

Group-by

Array Databases beat Relational Database tables
on storage efficiency & array computations

Relational Database

I	J	value
0	0	32.5
1	0	90.9
2	0	42.1
3	0	96.7
0	1	46.3
1	1	35.4
2	1	35.7
3	1	41.3
0	2	81.7
1	2	35.9
2	2	35.3
3	2	89.9
0	3	53.6
1	3	86.3
2	3	45.9
3	3	27.6

48 cells

Array Database

32.5	46.3	81.7	53.6
90.9	35.4	35.9	86.3
42.1	35.7	35.3	45.9
96.7	41.3	89.9	27.6

16 cells

- Math functions run directly on native storage format
- Dramatic storage efficiencies as # of dimensions & attributes grows
- High performance on both sparse and dense data

Status and Performance

- SciDB is 100x Postgres on analytics
- SciDB is faster or the same on vanilla data management
- SciDB is comparable to R on analytics
  - But scales!

Broad range of early adopters

Commercial

- Major pharma company
- Major insurance company
- Pricing analytics company

Scientific

- NCBI One Thousand Genomes project
- Lawrence Berkeley National Labs
- NASA Goddard

SciDB-R

- R user interface
- SciDB array objects are not limited by standard R array indexing limits
- Scrape off big array manipulation and send to SciDB

Summary

- As the world moves from simple analytics to complex ones
  - RDBMS likely to fail
  - And Hadoop unlikely to scale
- Check out SciDB
  - Download from www.scidb.org