MAD SKILLS
NEW ANALYSIS PRACTICES FOR BIG DATA

JEFF COHEN
BRIAN DOLAN
MARK DUNLAP
JOE HELLERSTEIN
CALEB WELTON

GREENPLUM
FOX AUDIENCE NETWORK
EVERGREEN TECHNOLOGIES
UC BERKELEY
GREENPLUM
MADGENDA

- Warehousing and the New Practitioners
- Getting MAD
- A Taste of Some Data-Parallel Statistics
- Engine Design Priorities
IN THE DAYS OF KINGS AND PRIESTS

- Computers and Data: Crown Jewels
- Executives depend on computers
  - But cannot work with them directly
- The DBA “Priesthood”
  - And their Acronymia
    - EDW, BI, OLAP
THE ARCHITECTED EDW

- Rational behavior … for a bygone era

  “There is no point in bringing data … into the data warehouse environment without integrating it.”
  — Bill Inmon, Building the Data Warehouse, 2005
NEW REALITIES

- TB disks < $100
- Everything is data
- Rise of data-driven culture
  - Very publicly espoused by Google, Wired, etc.
  - Sloan Digital Sky Survey, Terraserver, etc.
The quest for knowledge used to begin with grand theories.

Now it begins with massive amounts of data.

Welcome to the Petabyte Age.
MAD SKILLS

- **Magnetic**: 
  - attract data and practitioners

- **Agile**: 
  - rapid iteration: ingest, analyze, productionalize

- **Deep**: 
  - sophisticated analytics in Big Data
MAD SKILLS

**mad skills**

To be able to do/perform amazing/unexpected things

*I gots me mad skills, yo.*

To be said after performing an extraordinary feat.
“Looking for a career where your services will be in high demand?

... Provide a scarce, complementary service to something that is getting ubiquitous and cheap.

So what’s ubiquitous and cheap?
Data.

And what is complementary to data?
Analysis.

Hal Varian, UC Berkeley, Chief Economist @ Google
THE NEW PRACTITIONERS

- Aggressively Datavorous
- Statistically savvy
- Diverse in training, tools
• Greenplum DB
  • 42 Sun X4500s ("Thumper") each with:
    • 48 500GB drives
    • 16GB RAM
    • 2 dual-core Opterons

• Big and growing
  • 200 TB data (mirrored)
  • Fact table of 1.5 trillion rows
  • Growing 5TB per day
  • 4-7 Billion rows per day

• Variety of data
  • Ad logs, CRM, User data

• Research & Reporting
  • Diversity of users from Sales Acct Mgrs to Research Scientists
  • Microstrategy to command-line SQL

• Also extensive use of R and Hadoop

As reported by FAN, Feb, 2009
MADGENDA

- Warehousing and the New Practitioners
- Getting MAD
- A Taste of Some Data-Parallel Statistics
- Engine Design Priorities
VIRTUOUS CYCLE OF ANALYTICS

- Analysts trump DBAs
  - They are data magnets
  - They tolerate and clean dirty data
  - They like *all* the data (no samples/extracts)
  - They *produce* data

Figure 1: A Healthy Organization
MADGENDA

- Warehousing and the New Practitioners
- Getting MAD
- A Taste of Some Data-Parallel Statistics
- Engine Design Priorities
A SCENARIO FROM FAN

How many female WWF fans under the age of 30 visited the Toyota community over the last 4 days and saw a Class A ad?

How are these people similar to those that visited Nissan?

Open-ended question about statistical densities (distributions)
- Data Mining focused on individual items
  - Statistical analysis needs more focus on *density* methods!

- Need to be able to utter statistical sentences
  - And run massively parallel, on Big Data!

1. (Scalar) Arithmetic
2. Vector Arithmetic
   - I.e. Linear Algebra
3. Functions
   - E.g. probability *densities*
4. Functionals
   - I.e. functions on functions
   - E.g., A/B testing: a *functional over densities*
5. Misc Statistical methods
   - E.g. resampling

May all your sequences converge
ANALYTICS IN SQL @ FAN

- Paper includes parallelizable, statistical SQL for
  - Linear algebra (vectors/matrices)
  - Ordinary Least Squares (multiple linear regression)
  - Conjugate Gradient (iterative optimization, e.g. for SVM classifiers)
  - Functionals including Mann-Whitney U test, Log-likelihood ratios
  - Resampling techniques, e.g. bootstrapping

- Encapsulated as stored procedures or UDFs
  - Significantly enhance the vocabulary of the DBMS!
- These are examples.
  - Related stuff in NIPS ’06, using MapReduce syntax
  - Plenty of research to do here!!
MADGENDA

- Warehouse and the New Practitioners
- Getting MAD
- A Taste of Some Data-Parallel Statistics
- Engine Design Priorities
PARALLELISM AND PLURALISM

- MAD scale and efficiency: achievable only via *parallelism*
- And *pluralism* for the new practitioners
  - Multilingual
  - Flexible storage
  - Commodity hardware
- Greenplum a leader in both dimensions
ANOTHER EXAMPLE

- Greenplum DB, 96 nodes
  - 4.5 petabytes of storage
  - 6.5 Petabytes of user data
    - 70% compression
  - 17 trillion records
  - 150 billion new records/day

As reported by Curt Monash, dbms2.com. April, 2009
PLURALISTIC STORAGE IN GREENPLUM

- Internal storage
  1. Standard “heap” tables
  2. Greenplum “append-only” tables
     - Optimized for fast scans
     - Multiple levels of compression supported
  3. Column-oriented tables
  4. *Partitioned* tables: combinations of the above storage types.

- External data sources
SG STREAMING

- Parallel many-to-many loading architecture
  - Automatic repartitioning of data from external sources
  - Performance scales with number of nodes
- Negligible impact on concurrent database operations
- Transformation in flight using SQL or other languages
- 4 Tb/hour on FAN production system
MULTILINGUAL DEVELOPMENT

- SQL or MapReduce
- Sequential code in a variety of languages
  - Perl
  - Python
  - Java
  - R
- Mix and Match!
• Unified execution of SQL, MapReduce on a common parallel execution engine

• Analyze structured or unstructured data, inside or outside the database

• Scale out parallelism on commodity hardware
I Got MAD Skills

THE NEW WAY OF DATA WAREHOUSING
TIME FOR ONE?
BOOTSTRAPPING

- A Resampling technique:
  - sample \( k \) out of \( N \) items with replacement
  - compute an aggregate statistic \( \theta_0 \)
  - resample another \( k \) items (with replacement)
  - compute an aggregate statistic \( \theta_1 \)
  - ... repeat for \( t \) trials

- The resulting set of \( \theta_i \)'s is normally distributed
  - The mean \( \theta^* \) is a good approximation of \( \theta \)
  - Avoids overfitting:
    - Good for small groups of data, or for masking outliers
Tricks:

- Given: dense row_IDs on the table to be sampled
- Identify all data to be sampled during bootstrapping:
  - The view `Design(trial_id, row_id)` easy to construct using SQL functions
  - Join `Design` to the table to be sampled
    - Group by `trial_id` and compute estimate
    - All resampling steps performed in one parallel query!
- Estimator is an aggregation query over the join
- A dozen lines of SQL, parallelizes beautifully
1. CREATE VIEW design AS
   SELECT a.trial_id, floor (N * random()) AS row_id
   FROM generate_series(1,t) AS a (trial_id),
   generate_series(1,k) AS b (subsample_id);

2. CREATE VIEW trials AS
   SELECT d.trial_id, theta(a.values) AS avg_value
   FROM design d, T
   WHERE d.row_id = T.row_id GROUP BY d.trial_id;

3. SELECT AVG(avg_value), STDDEV(avg_value)
   FROM trials;