MAD Skills: New Analysis Practices for Big Data
• Centralized Data Warehouses with well architectured data and functions in a ETL fashion (Extract, Transform and Load)
• Provides for basic operations on the database, accessible through regular business intelligence tools
• What is the problem with this approach?
MAD analytics

• Magnetic - Attract data and practitioners. The database must be painless and efficient

• Agile - Analysts need to ingest and analyze in the fly

• Deep – Sophisticated analytics at scale. Don’t force analyst to work on samples (where the long tail is lost)
Why MAD for IOT

Sensors

Aggregators

Big Data + Machine Learning

Mobile App/Visualization
Datawarehousing for MAD

• Should provide for statistical computation and running of ML algorithms on the database
• Allow for the parallelizing of the complex queries run on the database.
• Should allow for its physical and logical contents to be in continuous rapid evolution
• Need to provide a system which people from diverse background can use
Map Reduce and Parallel Programming

• Programming model in extensive use for performing complex statistical operations
• Has caused a number of statistically-minded researchers and developers to focus on big data and data-parallel programming
• Extensive work has been done to implement ML algorithms on the map reduce framework
• Hence, influencing the design of the data warehousing techniques employed for MAD analytics
Greenplum Database System

- Was used in MAD analytics for Fox Audience Network
- Run on 42 nodes (40 nodes for query processing and two master nodes - one for failover)
- Sun X4500 Thumper was used for query processing.
- 16GB ram
- 48 500GB drivers
- 1 table 1.5 million rows
- Many different types of user workloads
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)

Dolon, Brian. [Photograph]. Retrieved from presentation on MAD skills.
How does Greenplum work?

- Traditional SQL
- Vector arithmetic
- Functional arithmetic
Vectors and Matrices

Matrix is stored as a relation with schema \((\text{row\_number} \text{ integer}, \text{vector} \text{ numeric []})\)

How would we implement matrix addition?

- SELECT \text{A.row\_number, A.vector} + \text{B.vector} \text{ FROM A, B WHERE A.row\_number = B.row\_number;}
- Requires implementing vector addition

What about multiplication of a matrix and a vector?

- SELECT 1, \text{array\_accum(row\_number, vector*v)} \text{ FROM A;}
- Requires implementing dot product calculation of vectors
Vectors and Matrices

• What about multiplication of a matrix and a vector?

\[
\begin{bmatrix}
  a & b & c \\
  d & e & f \\
  g & h & i \\
\end{bmatrix}
\begin{bmatrix}
  x \\
  y \\
  z \\
\end{bmatrix}
= 
\begin{bmatrix}
  ax + by + cz \\
  dx + ey + fz \\
  gx + hy + iz \\
\end{bmatrix}
\]

– `SELECT 1, array_accum(row_number, vector*v) FROM A`
– Requires implementing dot product calculation of vectors
Vectors and Matrices

• What about matrix multiplication?

\[
\begin{bmatrix}
    a & b \\
    c & d \\
\end{bmatrix} \times \begin{bmatrix}
    e & f \\
    g & h \\
\end{bmatrix} = \begin{bmatrix}
    ae + bg & af + bh \\
    ce + dg & cf + dh \\
\end{bmatrix}
\]

A, B and C are square matrices of size $N \times N$

a, b, c and d are submatrices of A, of size $N/2 \times N/2$

e, f, g and h are submatrices of B, of size $N/2 \times N/2$

• Will the current representation do? What if the matrix is sparse?

• We use a sparse representation of the form (row_number, column_number, value)

• SELECT A.row_number, B.column_number, SUM(A.value * B.value) FROM A, B WHERE A.row_number = B.column_number GROUP BY A.row_number, B.column_number
Vectors and Matrices

• Few more things provided as part of vector arithmetic
  – TF-IDF and Cosine Similarity
  – Ordinary Least Squares
Traditional SQL provides basic statistics such as mean, variance etc.

Greenplum provides data-parallel implementations of comparative statistics such as Z-score, Log-Likelihood Ratios, Resampling techniques as primitives through functionals.
Mann-Whitney U Test

• How does this work
  – Assign numeric ranks to all the observations. $R_1$ is the sum of all the ranks.
  – $U = R_1 - (n_1(n_1 + 1) / 2)$, where $n_1$ is the sample size for sample 1, and $R_1$ is the sum of the ranks in sample 1
• This is implemented as a procedure called *mann whitney* using SQL queries and can be used by an analyst via a simple `SELECT mann whitney(value) FROM table` call
• In the same manner log-likelihood ratios and more procedures are implemented
MAD DBMS

- Loading and Query Processing
  - Scatter / Gather Streaming - fast

- Eliminates overhead of landing data and keeping it refreshed through the use of external tables

- Provides map reduce support
Storage

• Tunable table types:
  – external tables (e.g. files)
  – heap tables (frequent tables)
  – append-only tables (rare updates)
  – column-stores flexibility

• All tables have a distribution policy - achieved through partitioning
Conclusion

• Mad Programming allows for *map* and *reduce* functions can be written in Python, Perl, or R
• Future
  – Automatically choose data and tables layouts
  – Package management to enable better code reuse