Introduction

Treeminer develops and markets patented technology to dramatically speed up data mining tasks on large datasets.

Multi-billion dollar market
Growing 10%+ per year

- Proven Technology
- Four Patents Issued
- Experienced leadership
The Problem: Current Data Analytics Methods Stressed by Massive Data Growth

Data doubles every two years

Server performance doubles every four years

Algorithm execution time grows exponentially as datasets grow

"Extracting Value from Chaos", IDC, June, 2011

"How Much Information 2010", Short, Bohn, Baru, January, 2011
Today’s Answers: Tactical Responses

**Look at Less Data**

- “GOOD”

**Add Servers**

- “BETTER”

**but…**

- “BEST”

---

- What about Accuracy?
- What about Cost?

---

*Can we address the “Gap” algorithmically?*
Insights to Action – Drivers, Enablers, and Blockers

Unprecedented growth in data

Challenges to Knowledge Extraction

- Velocity
- Variety
- Volume

New Database Paradigms

New Storage and Execute Paradigms

New Compute Paradigms
The Dimensions of the Problem

- **Velocity**
- **Volume**
- **Variety**
- **Implementation**
- **Cost**

**Hardware**
- In Memory
- Data Scientist
- Clusters
- Batch
- Real Time

**Cost**
- Bio
- Log
- Structured

**Volume**
- PB
- TB

**Variety**
- Text

**Implementation**
- Hardware
- In Memory
- Data Scientist
- Clusters
- Batch
- Real Time

**Cost**
- In Memory
- Data Scientist
- Clusters
- Batch
- Real Time

**Velocity**
- Hardware
- In Memory
- Data Scientist
- Clusters
- Batch
- Real Time

**Volume**
- PB
- TB

**Variety**
- Text
Scalable Interoperable Solution Sets
Unified Architecture for Big Data Mining

Big Data Paradigm Agnostic

- Standalone Linux Environment
  - Available
- Clustered Big Data Environment (Hadoop)
- Streaming Data Mining classifies as data arrives
  - Beta Summer
- In Memory (Spark, Hama)

Algorithms scale from single process through to streaming Big Data Solutions
Our Solution

**Dramatically Improves Data Analytics Performance – Hours to Minutes**

- Integrates with existing big data – Hadoop Map/Reduce
- Organizes data for knowledge discovery
- Improves **Accuracy**

Software
Organize Data for Knowledge...

TODAY

- Horizontal Rows of Data
- Vertical Strips of Data (pTree)

<table>
<thead>
<tr>
<th>Row by row processing</th>
<th>Bit column by bit column processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured or unstructured</td>
<td>Structured or unstructured</td>
</tr>
<tr>
<td>Rows grow as dataset grows</td>
<td>Columns stay constant as dataset grows</td>
</tr>
<tr>
<td>Complex kernel or distance functions drive computational load</td>
<td>Simple logical operations (AND/OR) implement complex algebra to keep it computationally light</td>
</tr>
</tbody>
</table>
Data Mining in Hadoop - Challenges

• Current approaches:
  • Connectors (most Data Mining vendors)
    • Import and Export data from Hadoop file system to proprietary, non Map/Reduce applications
    • Lots of data movement, lack of ability to benefit from parallel framework
    • Avoid difficulty of making algorithms work
  • Open source frameworks (R, Mahout)
    • Widespread usage
    • Not optimized for Map/Reduce, better algorithms not widely available due to challenges
  • Emerging Frameworks
Platform for Big Data

- Parallel, low cost clustering environment for Big Data
  - Parallel execution framework (Map/Reduce)
- Rapidly emerging as industry standard – with YARN will be baseline for all big data clusters
Streaming Analytics

- Shared compute model with Hadoop
- Unified architecture enables off line model updates in Hadoop (with access to historical data stores) combined with real time views
- Ability to ask “different questions” of same streaming data
Treeminer Advantages

- Vertical framework offers significant advantages
  - New capabilities by integrating clustering and classification
  - Scalability – only approach with dimensional and cardinal scalability
  - Non-iterative predictive classification approach reduces infrastructure expense and produces results faster
  - Data Variety
- Flexible deployment paradigm matches requirement to infrastructure
Existing Hadoop Approaches

- Iterative approach not efficient in Map / Reduce
- Input “Splits” are fixed by rows of data, meaning all attributes get processed, needed or not
- Scalability limited by dataset
Treeminer’s Scalable Approach

- Single Pass
- Only relevant attributes are processed
- Vertical approach means extremely flexibility to maximize mapper resources
- Vertical algebra implemented in Mapper and Reducer means significant performance gains within Hadoop framework
Unified Approach

- Linux
  - Model updates
  - Test Datasets

- Storm
  - Access to same HDFS data
  - Streaming data arrival paradigm

- Hadoop Cluster (YARN)
  - Same algorithms
  - Same model
  - Unified access and execution
  - Interoperable Clustering / Classification framework

- Map / Reduce
  - Access to Historical Data
  - Batch Oriented (models, batch, etc)
Treeminer In-House Analytics

- Based on vertical data structures
- Predictive Classification
  - Oblique
  - SPS – Supervised anomaly detection / ranking
- Clustering
  - k-means
  - Auto-clustering (Summer)
- Interoperability
Performance Increases as Nodes Increase
Text Classification Performance

Growing document stores slow discovery with existing methods

Treeminer shows only small increases in classification time as document stores grow in size

In an iterative process (model – evaluate – produce), analytics times can be a critical factor in document production
Structured Data Performance

- Excellent scalability characteristics
- Significantly improved performance as compared to state of the art classification methods

Today's state of the art shows large increases in execution time as the dataset grows.

Treeminer's performance advantage grows as the dataset grows – orders of magnitude faster!
vMiner Platform

- Supported on Centos / Red Hat Linux 6.x and other Linux distributions
- Hadoop Cloudera distribution 3u5 and later, Storm 0.9.1
- Accumulo NO-SQL 1.4 and later
- SQL support (Oracle, MySQL, etc)
- Unstructured / structured data
- Beta Q1 2014
Recognition for Accuracy

Winner of prestigious 2006 ACM KDD Cup Data Mining Competition

- Utilized patented “meta classification” method
- Analysis of CAD data to identify images free of pulmonary embolism - accuracy based
- Scored twice as high as second place

Significant performer in Netflix competition
Improve prediction of user ratings of movies based upon past ratings
Summary

Current methods of scaling rely on big hardware or reduced accuracy.

Treeminer’s approach is *algorithmic* and more scalable, patented, and proven across most common data mining activities.

Companies fully leveraging all the data in their possession will outperform their peers.

Mark Silverman
msilverman@treeminer.com