Deep Data Exploration:

*Find Patterns in Your Data – Faster & Easier*

Curt Monash, Founder and President, Monash Research
Francois Ajenstat, Tableau
Stephanie McReynolds, Aster Data
Steve Wooledge, Aster Data

March 10th, 2011
Topics

• Investigative Analytics: New Techniques in Data Exploration

• Data Exploration in Action: Aster & Tableau Demo

• Q&A
Who Is Aster Data?

Founded 2005, the fastest growing company in ‘big data’ market

- **Aster Data:** An analytics platform that stores and analyzes relational and non-relational data
- **Foundation of Aster’s Solution:** Is a MPP row/column MPP DBMS with a built-in analytics engine
- **What’s unique about Aster Data?** Includes SQL plus the popular MapReduce analytics framework

Customers

![Customer Logos]

Data Visualization / BI Partners

![Data Visualization Logos]
On March 3rd, Teradata announced its intent to acquire Aster Data to enable a new class of big data management and analytics.

**Teradata + Aster Data**

- **EDW**
  - OLAP and Reporting \(\rightarrow\) Data Exploration
- **Big Data**
  - Structured, Relational Data \(\rightarrow\) Diverse Data

**User Groups**
- Executives
- Business Users
- Analysts
- Data Scientists
Deeper Analytics is a Competitive Advantage

<table>
<thead>
<tr>
<th>Internet Use Cases</th>
<th>Financial Services Use Cases</th>
<th>Retail Use Cases</th>
<th>Media &amp; Information Services Use Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Social networking graph analysis</td>
<td>• Bottom-up, user behavior based risk analysis</td>
<td>• Cross-channel marketing attribution</td>
<td>• Predictive and granular forecasting</td>
</tr>
<tr>
<td>• Crowd-sourcing</td>
<td>• Real-time fraud &amp; link analysis</td>
<td>• Merchandising optimization</td>
<td>• Click-stream analysis</td>
</tr>
<tr>
<td>• Virality analysis</td>
<td>• Tick data analysis</td>
<td>• Consumer buying patterns / behavior</td>
<td>• Consumer segmentation &amp; micro-targeting</td>
</tr>
<tr>
<td>• Content targeting</td>
<td>• Trading surveillance</td>
<td>• Seasonality Analytics</td>
<td>• Ad optimization</td>
</tr>
<tr>
<td>• Advanced click stream analysis</td>
<td>• Attrition prediction via behavior pattern matching</td>
<td>• Personalization/Recommendations</td>
<td></td>
</tr>
</tbody>
</table>

"Analytics themselves don't constitute a strategy, but using them to optimize a distinctive business capability certainly constitutes a strategy."

- T. Davenport & J. Harris, Competing on Analytics: The New Science of Winning
Why Customers Turn to Aster Data

1. Manage & Analyze Diverse Data Outside the EDW
   - Relational, non-relational data, network security analysis, click-stream data, machine data, CDR, location data, logs...

2. Interactive, Big Data Analytics
   - Ad hoc analytics + data exploration on big data
   - SQL and beyond, uses MapReduce and SQL-MapReduce™

3. Linear, Elastic Scaling on Commodity Hardware
   - 90% lower hardware costs than traditional DBMS/EDW
   - Elastic scaling – easy expansion & contraction

Aster Data + Tableau – enabling deeper, richer analytics
Polling Question: Pick the Biggest Frustration with Data Exploration

- Data extraction and preparation from source systems
- Query performance
- Ease of use end-user tools
- Lack of analytical expressiveness in SQL or existing BI tools
- Other – please specify
- None – completely satisfied
Topics

• Introductions

• Data Exploration in Action: Aster & Tableau Demo

• Q&A
Curt Monash
President and Founder, Monash Research

• ~30 years as a leading software analyst
• Consultant to users, vendors, and investors alike
• Since 1990 he has owned and operated Monash Research – DBMS2 and others
• Currently a columnist for InformationWeek
• Previous work
  - #1 ranked stock analyst while at PaineWebber
  - Co-founded Evernet, Inc., a $40 million networking systems integrator
• Ph.D. in Mathematics (Game Theory) from Harvard University
Investigative Analytics

New techniques in data exploration

Curt A. Monash, Ph.D.
President, Monash Research
Editor, DBMS2

contact @monash.com
http://www.monash.com
http://www.DBMS2.com
Agenda

- Six aspects of analytic technology
- Investigative analytics
  - Uses
  - Tools
  - Pitfalls
Illumination? Or just support?
Six things you can do with analytic technology

- Make an **immediate decision**.
- **Plan** in support of future decisions.
- **Research, investigate, and analyze** in support of future decisions.
- **Monitor** what’s going on, to see when it necessary to decide, plan, or investigate.
- **Communicate**, to help other people and organizations do these same things.
- Provide **support**, in technology or data gathering, for one of the other functions.
Investigative analytics defined

Seeking patterns in data via techniques such as

- Statistics, data mining, machine learning, and/or predictive analytics.
- The more research-oriented aspects of business intelligence tools.
- Analogous technologies as applied to non-tabular data types such as text or graph.

where the patterns are previously unknown.

Source: http://www.dbms2.com/2011/03/03/investigative-analytics/
Investigative analytics attitude

- Data is your most important and differentiating business asset, or pretty close to it
- You have to keep getting new value out of data to keep competing
- New value revolves around new insights
Analytic progression

- Trends
- Correlations
- Decisions

Source: http://xkcd.com/552/
The three key drivers of investigative* analytics

- Marketing, especially personalized
- Problem (or anomaly) detection and diagnosis
- Optimization (or planning)

*Non-investigative analytics also has a pure communication/reporting component ... although the whole exercise is pretty pointless unless somebody is expected to - at least potentially - make a decision based on the reported information at some point.
Marketing

- Matching offers to (potential) customers
  - Characterizing/identifying customers/customer groups
  - Testing offers, where offers can be:
    - Ads/messages
    - Deals
    - Products & product characteristics

- Data is the key
  - Transactions
    - Loyalty cards
    - Credit cards
  - Communications
    - Direct
    - Social media
  - Weblogs, etc.
Detect and diagnose problems and anomalies

- Manufacturing (classic equipment-focused)
- Manufacturing (modern warranty-focused)
- Customer satisfaction (bad and good)
- Network operation
- Bad actors
  - Terror
  - Fraud
- Risk
- Scientific discovery (haystack needle, meet magnet)
Optimization and planning

- Inventory optimization
- Distribution planning
- Price/revenue maximization
- Algorithmic trading
- The risk analysis revolution
Two aspects of investigative analytics

- Monitoring and sifting data
  - Exciting because it’s Fast, Fast, Fast!!* ...
  - ... and has cool visuals
- Serious math
  - Geek supremacy

*See also “Big, Big, Big!!!”
Monitoring and sifting data

- Cool dashboards
- Drilldown and query from those cool dashboards
Serious math

- Statistics, which overlaps with ...
- ... machine learning
- Graph theory
- Monte Carlo simulation
- Maybe more?
Investigative analytics pitfalls

- The future may not be like the past
- Don’t ignore what you can’t measure
- Privacy
Topics

• Introductions

• Investigative Analytics: New Techniques in Data Exploration

• Q&A
A new kind of Business Intelligence

RAPID-FIRE BUSINESS INTELLIGENCE
Caption:
This is the sum of customer debit and credit card charges for the year 2010 in a single San Francisco zip code. Data is displayed by physical business location. Consumer spend is further broken down at each location by transaction type.

Abrupt changes in individual and global spending patterns are key inputs into Bank X's behavioral models.
Declining Monthly Restaurant Transactions in Zip Code 94114

The number of transactions by month at restaurants in the 94114 zip code indicates a troubling shift in customer behavior starting in June 2010. Transactions per month have been steadily declining. Yet, average transaction amounts are unaffected.

A standard SQL aggregate helps an analyst observe this decline, but does not answer for the analyst:

* What are the drivers of the trend?
* How to influence the drivers?
* What actions to take to reverse the trend?
The count of distinct users who stop using their credit cards, after an active period of use. (Other services, like ATM usage and Direct Deposit services continue.) Line color is proportional to income, and line width is proportional to previous number of transactions.

Analysts can mine granular data efficiently for ad-hoc patterns like this using the Aster Data SQL-MapReduce framework and the pre-packaged nPath SQL-MapReduce function. The closest SQL equivalent would require multiple self-joins, rewriting the data set, the use of rank, lead and lag functions, and still would not find all card-switchers.

To take action on this newly identified behavioral pattern, the organization now needs to understand how to predict switching behavior.
Path Analysis Identifies Customers Who Switch Credit Card Providers

Using their credit cards, after an active period of use. (Other services, like 4G
consumer.) Line color is proportional to income, and line width is proportional to previous number of transactions per month.

Analysts can mine granular data efficiently for ad-hoc patterns like this using the Aster Data SQL-MapReduce framework. The closest SQL equivalent would require multiple self-joins, rewriting the data set, and still would not find all card-switchers.

To take action on this newly identified behavioral pattern, the organization now needs to understand how to prevent...
Path Analysis Identifies Customers Who Are Likely to Switch Credit Cards

The following SQL query is used to identify customers who are likely to switch credit cards:

```
SELECT
    transaction_type = 'Total Payment Due' and transaction_amount >= 100 as T,
    transaction_type = 'CC PAYMENT' and transaction_amount >= 100 as P,
    count(*) as total_transactions,
    count(*) as total_spending_months,
    count(*) as total_missed_payments
FROM
    npath(
        ON tran_94114
        PARTITION BY customer_id
        ORDER BY transaction_post_date
        PATTERN('T.0*.P.0*.+(NT.0*.NP.0*.+)*)
        MODE(NONOVERLAPPING)
        SYMBOLS('T' as T, 'P' as P, 'NP' as NP, 'NT' as NT)
    )
RESULT(
    first(debugging of T) as customer_type,
    first(customer_id of T) AS customer_id,
    first(customer_income of T) as customer_income,
    count(*) of O as total_transactions,
    count(*) of P as total_spending_months,
    count(*) of NP as total_missed_payments,
    count(*) of NT as total_nothing_charged,
    first(transaction_post_date of NT) AS date_of_first_low_charge,
```

This query is designed to identify customers who have a large number of transactions with high spending months and missed payments, indicating a high likelihood of switching credit cards.
# Restaurant Transactions Reveal Social Links Between Customers

Influence analysis is a powerful new analytic technique based on graph analysis which can predict customer churn behaviors.

A social link structure is built by identifying shared behaviors within customer transactions. Here we have used split-checks paid at restaurant locations in San Francisco to identify pairs of customers. Customers are ranked by the number of their split-checks.

Here analysts leverage SQL-MapReduce in ways not even possible to consider in SQL. For this analysis transaction data has been mined for the link structure, explicitly matching on date and restaurant location and executing a fuzzy match on charge post time and dollar amount spent.

Selecting one of these pairs and right clicking allows us to look at additional shared details between paired individuals.
Influence analysis is a powerful new analytic technique based on graph analysis which can predict customer churn behaviors.

A social link structure is built by identifying shared behaviors within customer transactions. Here we have used split-checks paid at restaurant locations in San Francisco to identify pairs of customers. Customers are ranked by the number of their split-checks.

Here analysts leverage SQL MapReduce in ways not even possible to consider in SQL. For this analysis, transaction data has been mined for the link structure, explicitly matching on date and restaurant location and executing a fuzzy match on charge post time and dollar amount spent.

Selecting one of these pairs and right clicking allows us to look at additional shared details between paired individuals.
Social Connections Influence Card Switching Behavior Over Time

Card switching follows a pattern whereby one card switcher often quickly influences those in his social network to also switch credit cards. The higher the number of shared transactions between social pairs, the more likely that one card switcher in the pair will influence the other card holder to switch. It is also likely that the higher the number of social connections an individual has, the more likely they themselves will be influenced to switch.
Card Holders Segmented by the Number of Social Links in Their Transactions

Caption
Predicting the probability of a card user switching cards becomes easier when their social links are known.

Here we see the population of consumers in our single zip code, segmented by their number of social connections.

By comparing the segmented populations and their switching behavior, we easily see that customers with at least two social connections are more likely to switch cards.

Customers with at least 2 social connections are 5-10x more likely to switch cards. In fact, each social link a customer has who has switched cards (red), increases the probability that the customer will himself switch cards (green).

Selecting the number above any pie chart and right clicking, allows you to obtain more demographic information about that customer segment in order to plan your retention strategy.
Card Holders Segmented by the Number of Social Links in Their Transactions

<table>
<thead>
<tr>
<th>Number of Social Links</th>
<th>Row Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>197</td>
</tr>
<tr>
<td>1</td>
<td>322</td>
</tr>
<tr>
<td>2</td>
<td>381</td>
</tr>
<tr>
<td>3</td>
<td>302</td>
</tr>
<tr>
<td>4</td>
<td>246</td>
</tr>
<tr>
<td>5</td>
<td>102</td>
</tr>
<tr>
<td>6</td>
<td>107</td>
</tr>
<tr>
<td>7</td>
<td>36</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
</tr>
</tbody>
</table>

Caption

Predicting the probability of a card user switching cards becomes easier when their social links are known.

Here we see the population of consumers in our single zip code, segmented by their number of social connections.

By comparing the segmented populations and their switching behavior, we easily see that customers with at least two social connections are more likely to switch cards.

Customers with at least 2 social connections are 5-10x more likely to switch cards. In fact, each social link a customer has who has switched cards (red), increases the probability that the customer will himself switch cards (green).

Selecting the number above any pie chart and right clicking, allows you to obtain more demographic information about that customer segment in order to plan your retention strategy.
This bar chart shows the breakdown of income levels covered by the population selected in the Most Influential report. Such information is used by the marketing organization at the bank to refine their direct marketing messages to account holders most at risk for card switching.

Select any pie chart in the Most Influential chart to update the data on this visualization specifically for that population.
Measuring the Potential Business Impact of Social Influencer Analysis

Monitoring customers' social cliques regularly, to catch signals in the data that friends are defecting, can identify customers at risk before the business is lost.

This plot shows all known card switchers from the past year (red). This is compared to the monthly percentage of customers that could have been prevented from switching (blue), if only the bank had monitored early warning signals in their transaction data.

The data here proves that the social switcher model, built on a combination of SQL-MapReduce social link analysis and nPath, predicts with 30% or more accuracy a consumer’s likelihood to switch.

The blue customers are those that could have been prevented from leaving using the bank’s new social graph insight, delivering significant revenue loss prevention.
Topics

• Introductions

• Investigative Analytics: New Techniques in Data Exploration

• Data Exploration in Action: Aster & Tableau Demo
Thank You!

• Questions

• More information
  - www.dbms2.com ; www.monash.com
  - www.tableausoftware.com
  - www.asterdata.com

• Contact us
  - Curt Monash - contact@monash.com
  - Francois Ajenstat - fajenstat@tableausoftware.com
  - Steve Wooledge - steve@asterdata.com