COMPUTERWORLD

THE POWER OF BIG DATA SYMPOSIUM

6.26.12 | NYC

Optimizing **Big Data** for Real-Time Business Transformation

COMPUTERWORLD
THE POWER OF BIG DATA SYMPOSIUM

From Push to Pull: Using Big Data to Get To Know Your Customer

Debra S. Domeyer
CEO
Oversee.net

COMPUTERWORLD
THE POWER OF BIG DATA SYMPOSIUM

"Big Data" Extremity Across Three Measures

VOLUME

Human produced Information
1999 – 12 exabytes
2006 – 160 exabytes
2011 – **Estimate of 2.7 zetabytes** up 48% from 2011 IDC

VELOCITY

The digital content is predicted to double in only **18 months** and **every 18 months** thereafter. IDC

VARIETY

80% of enterprise data will be unstructured, making integration very expensive Gartner Group

3 Key "Big Data" Activites:

Store…Process…Query

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Favorite Big Data Definition

Quote of Tim O'Reilly brings it all home:

"Companies that have massive amounts of data without massive amounts of clue are going to be displaced by startups that have less data but more clue."

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Oversee.net "Big Data" Definition

- when the size of the data itself became part of the problem.
  - Mobile, Social, Profiling, Logs, Transactional – Pedabytes of data
- when the "unstructured" and "dispersed" data could not be processed using conventional methods
- as an essential component of a "test and learn" mind-set

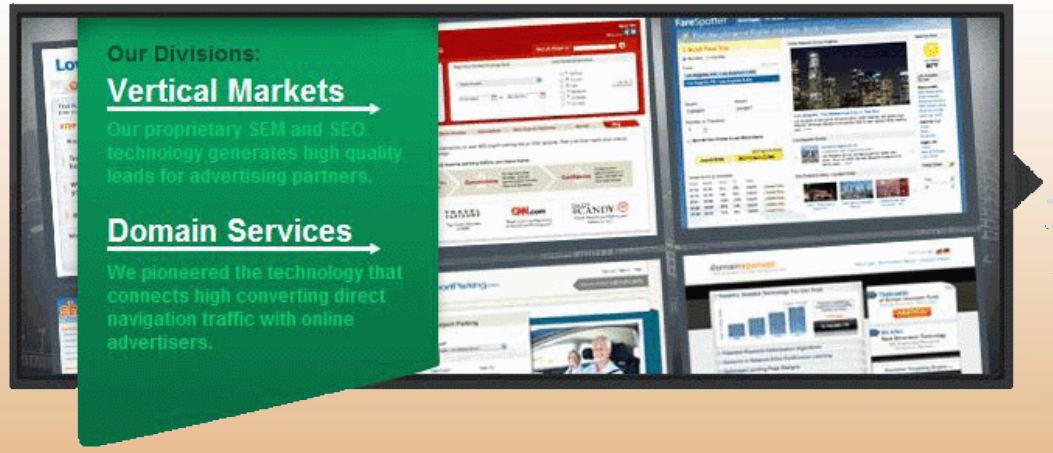
"The ability to rapidly test ideas fundamentally changes the company's mindset and approach to innovation"

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Who is Oversee.net?

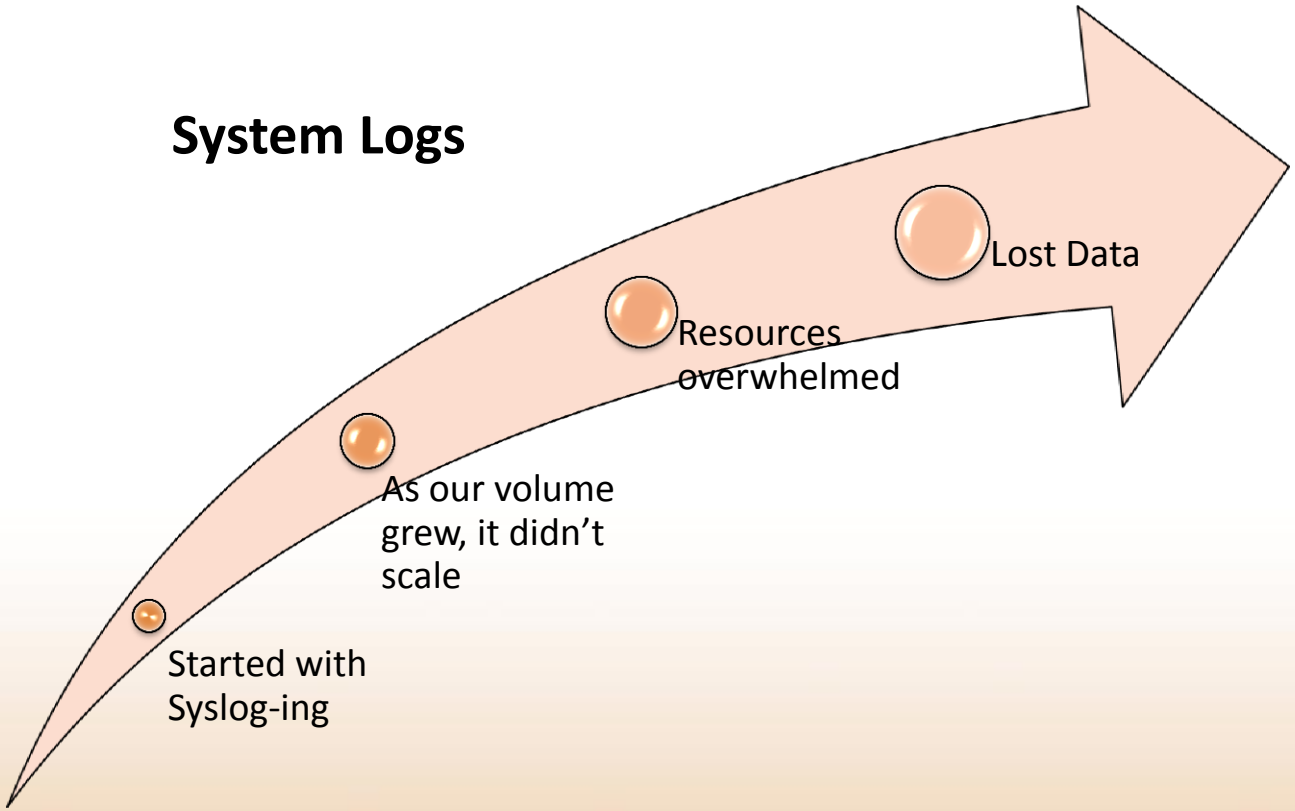
A Leader in Online Performance Marketing and Consumer Websites

- Owns more than 600,000 domain names; monetizes over 8 million domain names
- Large PPC and CPV network: providing advertisers with highly targeted access to 250 million unique visitors
- Premium Consumer Sites in Key Marketplace Segments
  (Travel, Retail, Consumer Finance)
- Over 14 million pre-qualified consumer leads/month funneled to advertising partners



COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Why the “Big Data” Evolution?



COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Why the "Big Data" Evolution?

Data is key differentiator in building our seasonal, temporal, emerging market business intelligence tools.

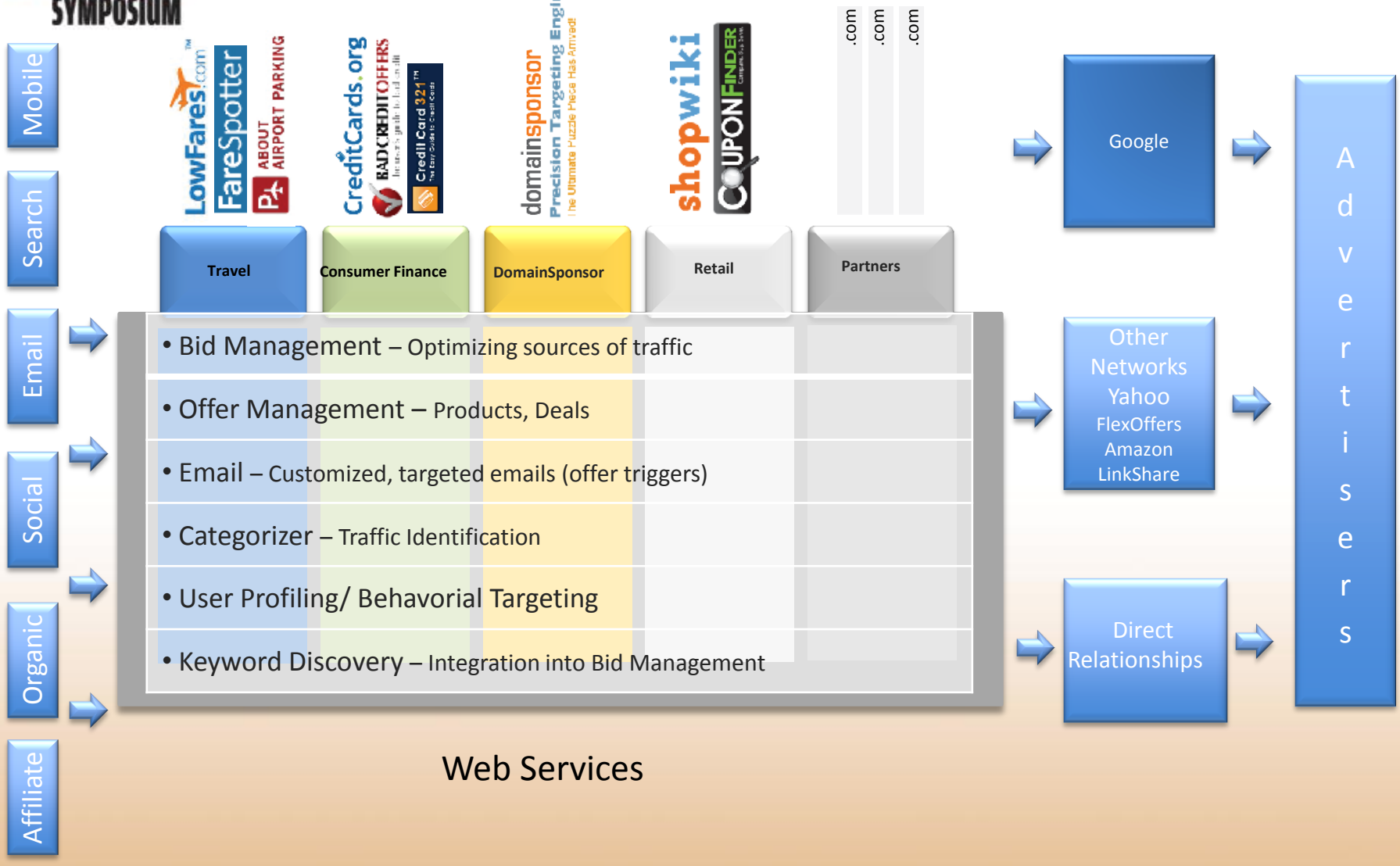
Data, Data, and More Data…

- 250 million Unique Visitors/Month
- 600 million Landing Pages/Month
- 50 million ad clicks/month
- 3,000 requests/second
- Daily raw data processing – 2 Terabytes
- Total data storage requirements – over 200 Terabytes

Process data daily: clicks, visits, searches, products, leads, emails, sessions, profiles…into petabytes of data!

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Performance Marketing Network



COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Data Analysis at Oversee.net

Data allows our publishers to better reach targeted audiences across platforms, countries, and devices

Categorizing across billions of keywords and millions of lander combinations allows us to make better lander selections.

Example: "Weight loss" is in the "Beauty" category, which does best on the "Gym" and "Scale" landers.

Category: Beauty & Personal Care **Order By:** searches

Keywords last updated: 12/5/2011 (month-to-date)

Keyword	Searches	Clicks	Revenue	RPS	RPC
lose weight	24,966	4,754	$3,786.62	$0.15	$0.80
weight loss program	21,897	6,034	$3,968.70	$0.18	$0.66
breast enlargement	21,258	9,887	$553.68	$0.03	$0.06
skin care	18,719	2,408	$1,708.49	$0.09	$0.71
weight loss	18,549	1,893	$1,220.60	$0.07	$0.64
skin care products	15,528	3,349	$2,621.05	$0.17	$0.78
makeup	15,259	3,714	$1,557.70	$0.10	$0.42

Landers last updated: 12/5/2011 (last 60 days)

Design ID	Design Name	RPS
741	Lifestyle-Gym	$0.26
909	Health-Weight Loss Scale	$0.25
786	Shopping - Jewelry	$0.16
717	Fragrances	$0.16
722	Finance-Family	$0.15
750	Kids-Monster	$0.15
761	Memphis-Pink	$0.15

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

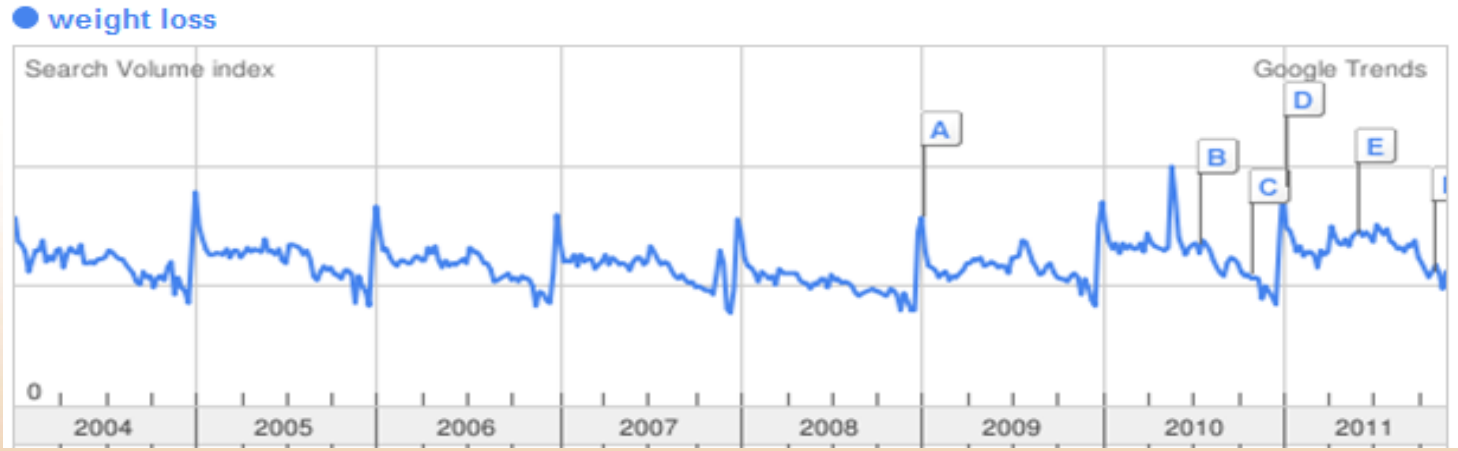
Data Analysis at Oversee.net

Keyword and Lander optimization

We can anticipate and promote seasonal keywords.

Example: “weight loss” is a popular search term in January

(New Year’s Resolution)

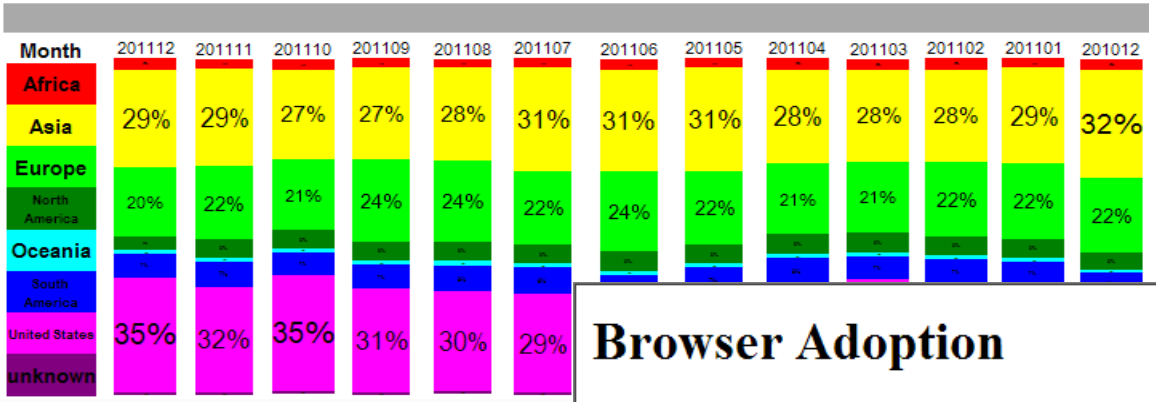


COMPUTERWORLD
THE POWER OF BIG DATA SYMPOSIUM

Business Intelligence Tools

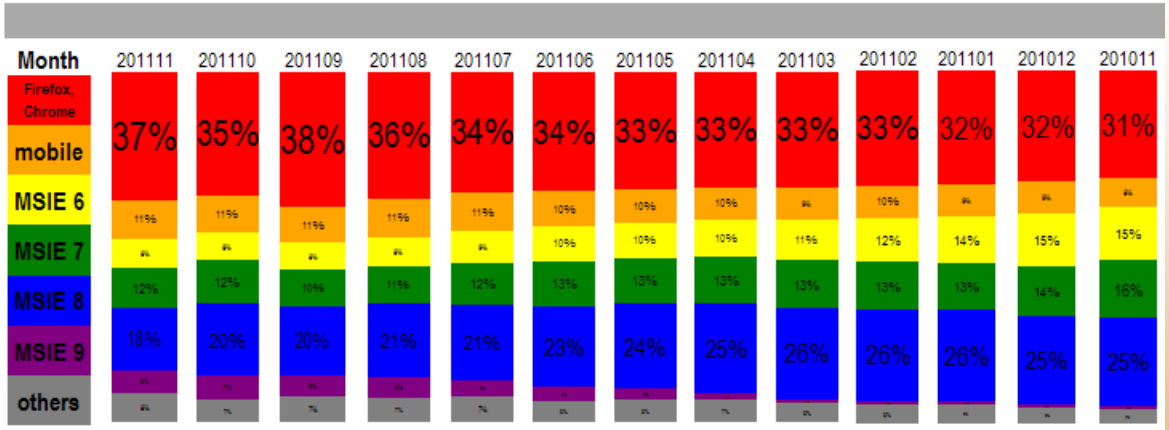
In an ever-changing Internet environment, we track and anticipate industry trends, such as shifts in browser use and the growth of mobile traffic.

Total Traffic by Country (unique loads)



Emerging markets overseas sparked an initiative to focus on the major countries of Europe.

Browser Adoption

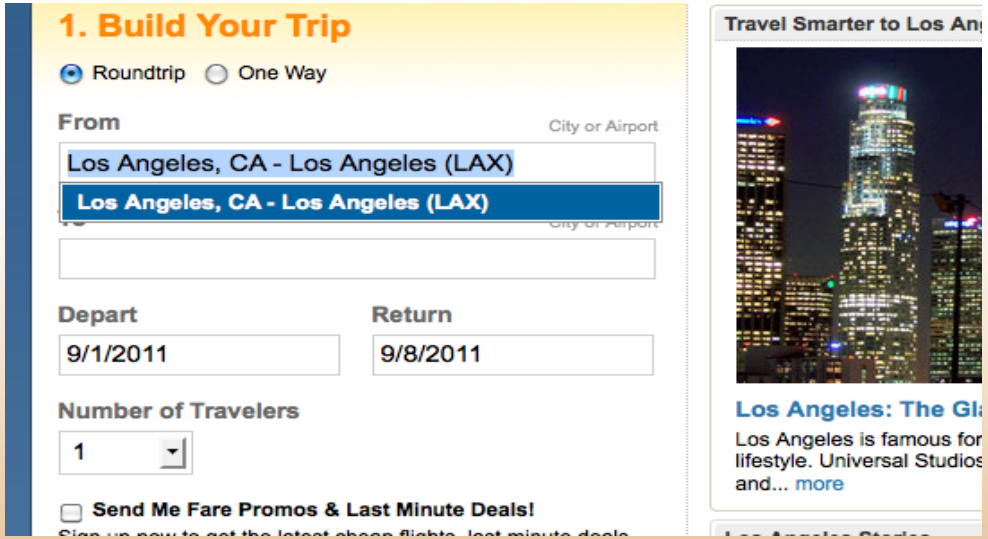


New browsers bring new features, which may impact the effectiveness of our product.

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

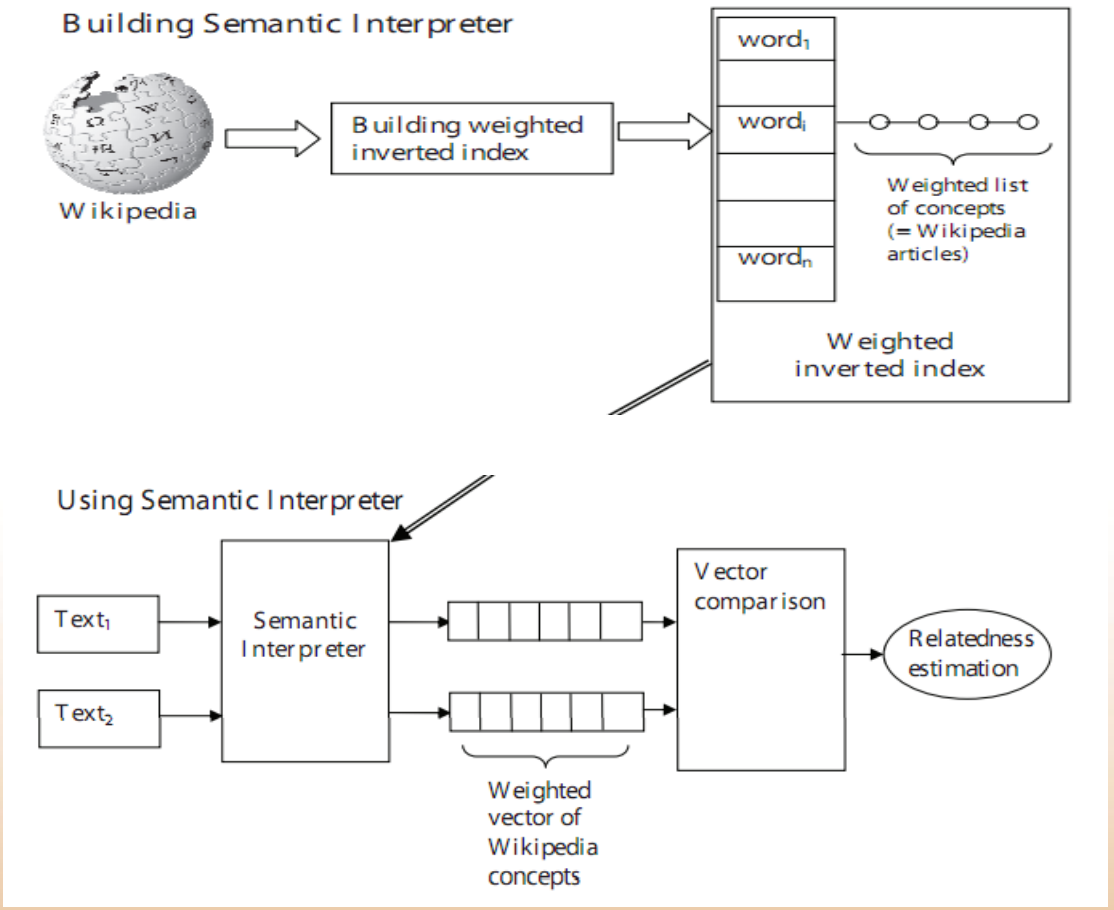
Mapping User Input to Locations using Explicit Semantic Analysis

- Develop location recommendation system based on semantic attributes
- System should be open-ended, i.e. impose minimum restrictions on user input
- Allow for system to update itself over time
- Be language independent



COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Inspiration: ESA



COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Mobile Technology

Proprietary Mobile Technology

Mobile Monetization
- Lander Optimizations
- Search & Mobile Ads
- Selling Mobile Traffic

Targeted Advertising
- Better targeting using Mobile Geo-location
- Mobile-Center Domain Buying

App Marketing
- Mobile App Marketing Opportunities
- Analytics and Measuring Metrics for Mobile & App Performance

Mobile Landers & Search

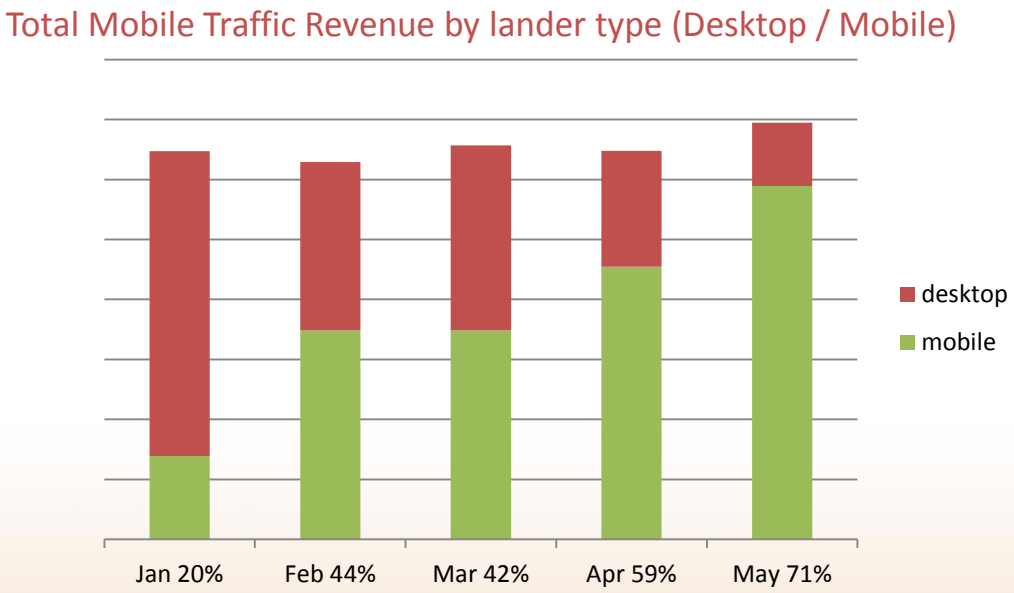
Mobile Ads & Geolocation

App Marketing

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Mobile Everywhere

Using Big Data, we identified Mobile traffic and began using landers that were specifically designed for Mobile users.



- In Jan, only 20% of Mobile traffic was on a lander designed for Mobile traffic.
- By May, 71% of Mobile traffic is now on Mobile landers.
- Revenue goes up 7% as we move Mobile traffic to Mobile landers.

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Data Types

Typical database teams, think of "structured data" only. Today's reality, forces us to deal with a lot of "unstructured data."

Unstructured Data	(Semi) Structured Data
• Web Crawling Textual Documents • Bitmap Data • Blog Posts • Emails • Images	• Web Logs • Data Models • XML Files • Software Code

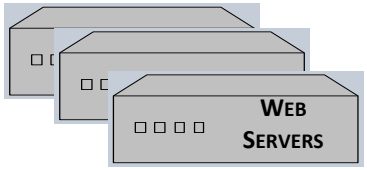
COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Netezza

- Multi-terabyte storage – Over forty TB's of storage capacity readily available for fast query response
- Our production system crunches on avg 92k queries/day
- Queries counting unique counts across multiple years can run from raw facts and return under 1 minute (no need to build multiple aggregate tables)
- Near real-time reporting due to fast processing window
- Our system supports EDW back-end processing and external online user support [can be achieved with Netezza]
- Most online queries (covering multiple months) are returned within seconds
- Very limited DBA/system support (less than 1 DBA for routine support)

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Data Flow Architecture



Internet Traffic

PERL LOG PROCESSING

"Turing" Scribe Mid Tier

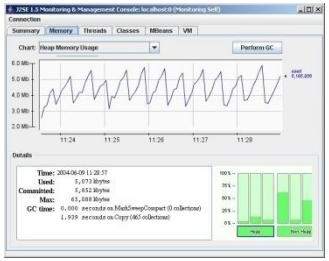
Initial Log Processing

MYSQL LEGACY DATABASE

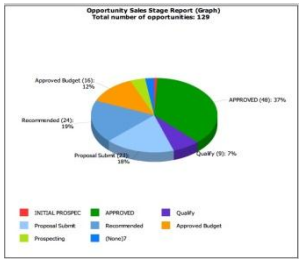
Stats DB

Additional Processing

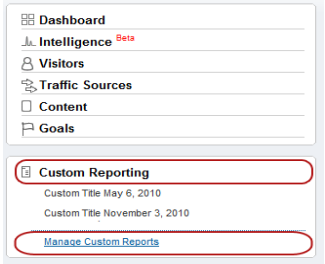
PUBLISHER REPORTING



SALES REPORTING TOOL



ANALYTICS DIRECT QUERYING



Application Access

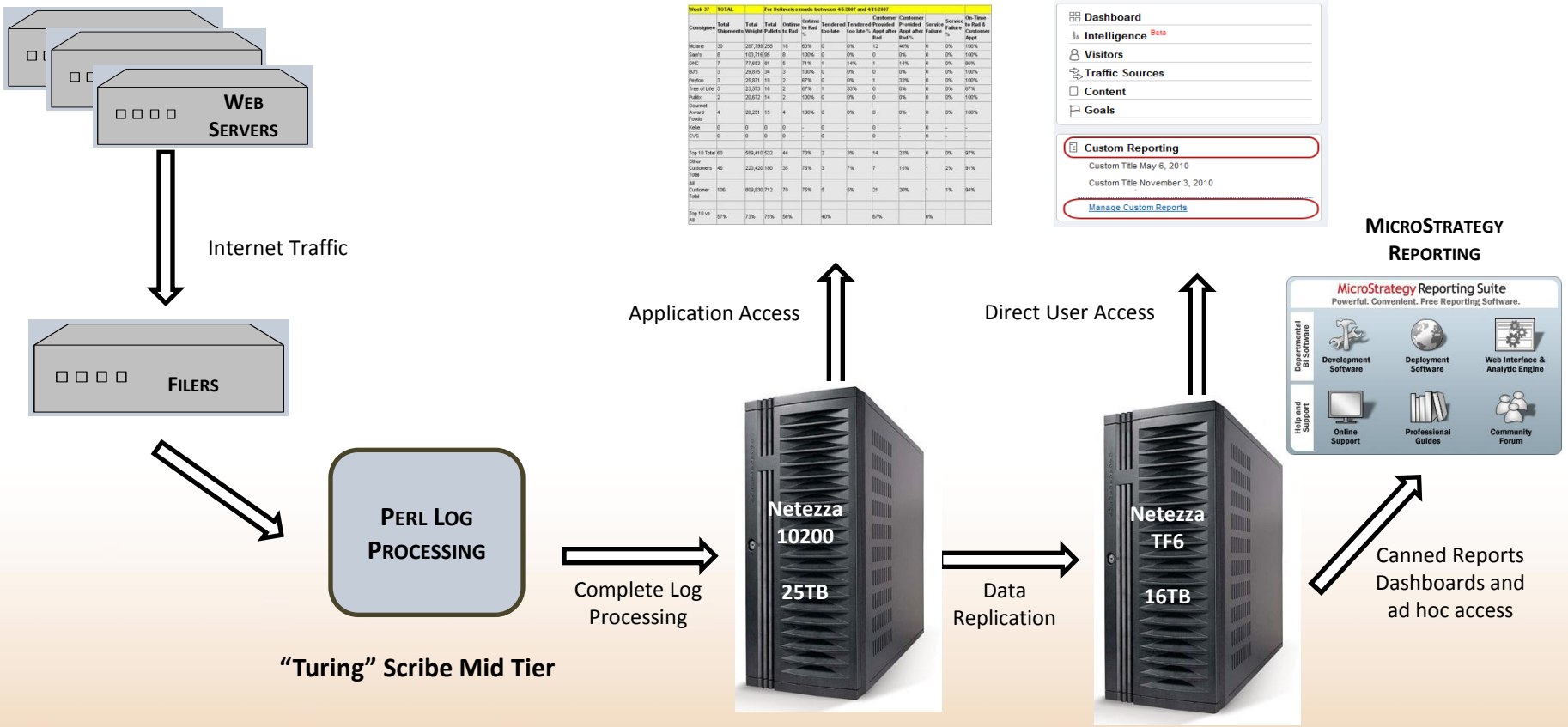
Direct User Access

Data Replication

- Dimensionalization Model for Reporting
- Additional Post Processing
- Aggregate Tables for Sub-second Publisher Stats

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

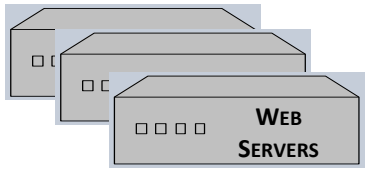
Data Flow Architecture



Remove legacy MySQL processing, introduce MicroStrategy BI tools.

COMPUTERWORLD
THE POWER OF BIG DATA SYMPOSIUM

Data Flow Architecture

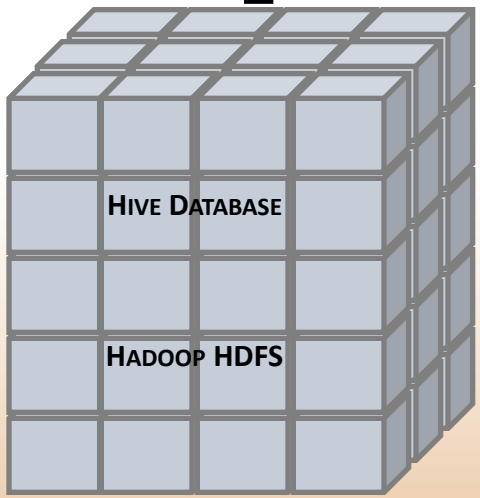


Internet Traffic

LOG COLLECTION AND PROCESSING

Log Processing

"Turing" Scribe Mid Tier



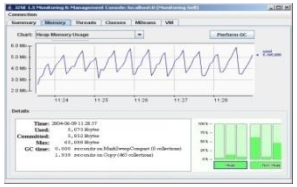
Hive and Pig Access

ADVANCED ANALYTICS

ANALYTICS

Direct Netezza Querying

PUBLISHER REPORTING



Canned Reports Dashboards and ad hoc access

MICROSTRATEGY REPORTING

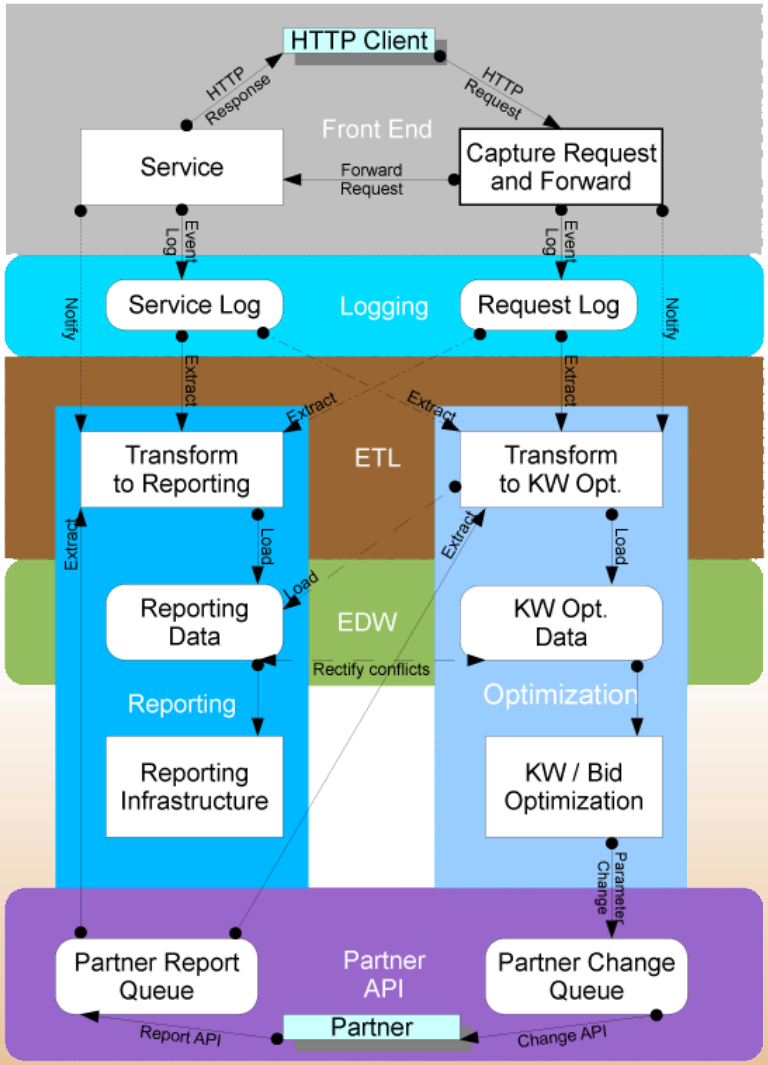
Application Access

Dimensional Data Migration

Introduce Hadoop cluster to handle future data volume growth (processing and strorage) and to support advanced analytics. Maintain Netezza appliances for fast querying.

COMPUTERWORLD
THE POWER OF BIG DATA SYMPOSIUM

A Single Example of Data Flow



COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Why Hadoop and Hive?

- Hadoop as Technical Opportunity
  - Store data as-is, in log files, etc., without first loading it into some database structure, and still be able to readily analyze it.
  - Parallelize without having to manage the details of parallelization
  - Can optionally use available, heterogeneous hardware

- Hadoop as Business Opportunity
  - Leverage data (especially historical data) by using more data than in just Netezza alone
  - Full user agent strings
  - Keyword impressions
  - Turing Internals
  - Conduct additional business analysis: Keyword impressions

- Hive
  - Simplifies managing and querying structured data built on top of Hadoop
  - Automates the Map Reduce code
  - HDFS for Storage
  - Metadata in an RDBMS
  - Allows Analysts easier access to data

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

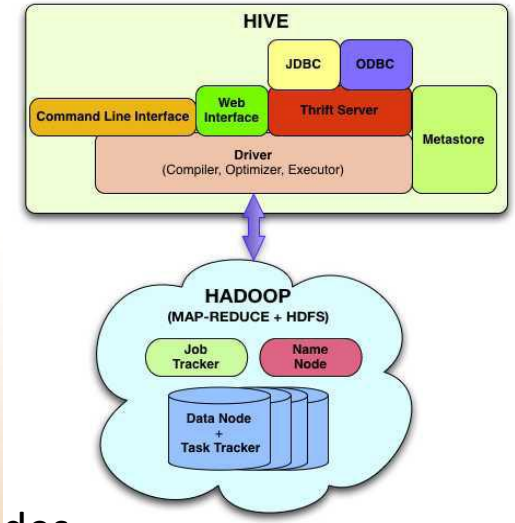
Why Hadoop and Hive?

- **Data Locality** - Hadoop tries to send the computation to where the data lives. It is easier to send a program than moving big data. This minimizes network calls and speeds up computation.
- **Redundancy** – Hadoop Data gets broken down into 64 MB blocks. Every block gets copied and gets distributed all over the file system (HDFS). Thus if a disk fails - the data is not lost. One can also duplicate jobs on the cluster. We can rely on other job's results if one fails.
- **Flexible API** - MapReduce API is quite simple to use and well abstracted. For instance, programmers can simply call read data() function without having to know what happens behind the scenes.
- **A fully fledged framework** - Large community support and has a "big" growing ecosystem – Hbase, Hive, Pig, Zookeeper, Avro, Chukwa…

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Hadoop / Hive Architecture

- Offers highly scalable and fault-tolerant processing of very large data sets.
- Map Reduce with Hadoop
- Runs on top of HDFS (Hadoop Distributed File System)
- Query predicate push down via server side scan and get filters
- Optimizations for real-time queries
- Supports row locks (built using ZooKeeper)
- A high performance Thrift gateway
- HTTP supports XML, Protobuf, and binary
- Cascading, hive, and pig source and sink modules
- jRuby-based (JIRB) shell
- No single point of failure
- Rolling restart for configuration changes and minor upgrades
- Random access performance is like MySQL



http://hadoop.apache.org/hive

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Hadoop Lessons Learned

- Get your team to use Pig and/or Hive whenever possible rather than implementing custom Map Reduce Jobs

  *If you think your job can be expressed in a relational query, pop it into Hive & see what happens

  *Even if you are going implement a custom Map Reduce Job think about organizing your data in a way that lends itself to being associated with a Hive/Howl/Pig metastore

- Learn how to use Hadoop counters!
  - They are your friend. Pop errors and other useful info into Hadoop counters and they'll show up aggregated in the Hadoop Web view

- At least try things out on the cloud
  - Avoid initial setup costs
  - Before investing in a Hadoop cluster, try out concepts and get performance metrics at scale quickly using cloud based services like Amazon's Elastic Map Reduce

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Hadoop Lessons Learned (cont.)

- Make sure rows are serialized using a backwards compatible format
  - Don’t get stuck having to think about column orders when adding and removing fields
  - Try using JSON, Thrift, Avro, etc.
  - SequenceFile, CSV, etc., are nice but the performance boost is negligble compared to the IO costs.
- Think about compression for the long-term.
- Consider streaming
  - If you have a team that's strong in another language like Perl/Ruby/Python/PHP/etc consider prototyping jobs using streaming when you need to create a custom job
- Don’t just throw Hadoop over the wall to the operations team
  - Operating a Hadoop cluster usually requires a fair amount of training.

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Integration Techniques

- Elephant Bird
  - Binary logs formatted as protocol-buffer messages. Elephant Bird a library that generates Hadoop and Hive bindings to read these messages.
- Mahout
  - Library of machine learning and data mining algorithms built on Hadoop.
  - Domain / User Categorization
  - Travel Recommendations
- Netezza Scoop and Microstrategy Connectors into Hadoop
- Cloudera – Providing support for implementation
- Map Reduce using Perl - the HadoopThriftServer class is part of the /usr/lib/hadoop/contrib/thriftfs/hadoop-thriftfs-0.20.2+737.jar0.20.2+737.jar file.
- Profiling - daily aggregates available for 60 days in raw form for processing to allow optimization stored in Hadoop in broken into fragments to allow for pattern recognition algorithms .

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

Safeguarding Your Data

- Multiple levels of redundancy across all systems (Netezza and Hadoop).
  - Backup mission critical data.
- NameNode – work is being done to safeguard NameNode. Losing NameNode means losing data.
- For the security of the Hadoop cluster you should encrypt the data using single-key block encryption and transport encryption for HDFS.

COMPUTERWORLD THE POWER OF BIG DATA SYMPOSIUM

What's Next?

Oversee.net's Data Initiatives

- Dwell time / repeat visit reporting at session and visitor level
- Full behavioral profiling integrated with demographics
- Full machine learning integration with both domain and keyword categorization
- Traffic channel integration with profiling
- Social profiling integrated to understand traffic
- Real-time profiling integrated with ad placement

COMPUTERWORLD
THE POWER OF BIG DATA SYMPOSIUM

Questions?

OVERSEE.NET

Debra Domeyer
Chief Executive Officer

E ddomeyer@oversee.net W www.oversee.net
T 213.408.0080
515 S. Flower Street, Suite 4400, Los Angeles, CA 90071

COMPUTERWORLD

THE POWER OF BIG DATA SYMPOSIUM

6.26.12 | NYC

Optimizing **Big Data** for Real-Time Business Transformation