TRENDS IN BIG DATA ANALYTICS: NEXT GENERATION ENTERPRISE DATA WAREHOUSING, AND PREDICTIVE ANALYTICS

Joe Shaffner, North America Technical Director
Sybase, an SAP company
DATA AND ANALYTICS – WHERE THE ACTION IS!

Data Warehouse Appliances

OLAP KXEN
Statistics
Multimedia Analytics
Univariate distributions
Principal component analysis
Zementis

Data Mining
Text Search

MapReduce

Columnar Database

Text Analytics
Predictive Model Markup Language

Hadoop

PMML Univariate Distributions

Linear regression

Monte Carlo simulation

Logistic regression

Fuzzy Logic

Naïve Bayes

Cluster analysis

Massively Parallel Processing

SYBASE® | An SAP Company
Information Armageddon: A Short History

1970: Alvin Toffler – “Information Overload”

1986: Theodore Roszak – “An excess of information may actually crowd out ideas ...”

1990: IEEE – “Crisis in Mass Storage”

2010: IDC – “By 2020 more data created than capacity to store...”

2011: McKinsey – “scale and scope ... inflection point”

BIG DATA, BIG OPPORTUNITY

10% CAN LEAD TO LARGE RETURNS

PRODUCTIVITY INCREASE
A 10% increase in **usability** of data translates to an increase of **$2.01 billion**
in total revenue per year.

SALES INCREASE
A 10% increase in **accessibility** to data translates to an additional
**$65.67 million** in net income per year.

Let's Look At Some Specific Industries

**PRODUCTIVITY INCREASES**

- Retail: 49%
- Consulting: 39%
- Air Transportation: 21%
- Construction: 20%
- Food Products: 20%
- Steel: 20%
- Automobile: 19%
- Industrial Instruments: 18%
- Publishing: 18%
- Telecommunications: 17%

**SALES INCREASES**

- Telecommunications: $9.6 bn
- Consulting: $5.0 bn
- Steel: $4.3 bn
- Automobile: $4.2 bn
- Food Products: $3.4 bn
- Air Transportation: $3.4 bn
- Construction: $2.0 bn
- Retail: $1.2 bn
- Industrial Instruments: $0.8 bn
- Publishing: $0.4 bn

Standout Industries In Sales Increases

- **TELECOMMUNICATIONS**
  - $9.6 billion = 114,825 OPERATIONS RESEARCH ANALYSTS
  - OR 64,000 TOWERS

- **AUTOMOBILE**
  - $4.2 billion = 48,275 MECHANICAL ENGINEERS
  - OR 175,000 PRIUSES

- **FOOD PRODUCTS**
  - $3.4 billion = 91,891 BUYERS/PURCHASING AGENTS
  - OR 809,523 CHILDREN FED IN AFRICA FOR TEN YEARS

- **AIR TRANSPORTATION**
  - $3.4 billion = 55,737 AIRFIELD OPERATION SPECIALISTS
  - OR 400 AIRPLANES

- **RETAIL**
  - $1.2 billion = 16,901 PURCHASING MANAGERS
  - OR 40,000 OM Commerce Apps

- **PUBLISHING**
  - $400 million = 6,577 EDITORS
  - OR 800,000 IPADS
“Companies adopting data driven analytics have 5% - 6% better top line and bottom-line.”
- McKinsey & Co

Median ROIs from Analytics:
- Operational Analytics: 277%
- Financial Analytics: 139%
- Customer Analytics: 55%
COMPETING ON DATA DRIVEN ANALYTICS
NEW “BIG DATA” FOCUS AREAS HAVE EMERGED

4 Factors
In Enterprise Analytics

- THE RELENTLESSNESS OF BIG DATA
  More data than storage capacity predicted

- THE USER POPULATION BOMB
  Millions of data miners in waiting

- UNCHARTED WATERS OF UNSTRUCTURED DATA
  Data variability becomes standard

- VELOCITY RAMPS UP DATA
  Data ingress kicks into gear

Marketing Analytics
New focus: digital channels
- Track visits, actions to find best channel
- Discover best mix: email, social media, search

Sales Analytics
New focus: correlation
- Deal DNA: correlate emails, chatter, meetings
- Predict deal risks based on pattern match

Operational Analytics
New focus: machine data
- Analyze weblogs, SMS, sensors - hunt down operational inefficiency
- Monitor, adjust, enforce risk mitigation

Financial Analytics
New focus: simulation
- Financial scenarios error margin simulation
- Analysis of portfolio returns, liquidity, ...

COMPETING ON DATA DRIVEN ANALYTICS
NEW “BIG DATA” FOCUS AREAS HAVE EMERGED
NEW FOCAL POINT FOR BIG DATA

BIG DATA ANALYTICS

The market is currently here

Big data analytics

Big data management

Market Adoption

The market for big data is evolving. Organizations were initially looking for big data management solutions that enable them to deal with the massive data volumes and variety... but the real value is derived from gaining insights in the data, which requires organizations to then look for big data analytics tools that provide insights.

Big Data Management
Primary focus: manage and federate the increasingly massive volume, variety and velocity of data?

Big Data Analytics
Primary focus: Derive business value and provide insight from all the data that is being federated and managed?
BIG DATA ANALYTICS GAP

DRIVING SIGNIFICANT INNOVATION IN THE DATABASE INDUSTRY
COMPETING ON SPEED OF ANALYSIS

REAL TIME ANALYTICS DRIVES VALUE

INTELLIGENCE VALUE

Surveillance Event

Data Integrated and Ready for Analysis

Information Delivered

Value Gained

Action Taken

T₀

T₁

Faster Time To Action

WHY THIS IS DIFFICULT

Complex Questions
Big Data
Usage Expansion

Analytics Requirements

Systems Scalability

Analytics Capability Gap

Projected Analytics Capabilities

V₀

V₁

V'
THE GREAT CIO CHALLENGE
HARMONIZE AND LEVERAGE RIGHT TECHNOLOGY & SKILLS

BUSINESS EXPECTATIONS

- Marketing Analytics
  New focus: digital channels

- Sales Analytics
  New focus: correlation

- Operational Analytics
  New focus: machine data

- Financial Analytics
  New focus: simulation

TECHNOLOGIES, SOLUTIONS, SKILLS

- Which technology should we consider?
  - Appliance
  - MapReduce
  - RDBMS
  - EDW
  - Column Store
  - In-DB Analytics

- Which skills set will we need?
  - BI Programmers
  - Data Scientists
  - Biz Analysts
  - DBAs
  - End Users

Is a unifying platform possible?
NEW DEMANDS NEED A NEW BUSINESS ANALYTICS PLATFORM

Traditional DBMSs
- Transactional DBs used for operational reporting
- Not intended for complex reporting and analytics

Data Warehouse & Data Mart DBMSs
- Designed to store massive volumes of complex data collected across the enterprise
- Accessed by a small number of users

Enterprise Data Warehouse 2.0?
- Designed to analyze massive volumes of complex data collected across the enterprise
- Designed for large numbers of users, workloads & interfaces
BIG DATA, WHAT IS IT? WIKIPEDIA SAYS >

- Not really what we used to call VLDB
- Datasets that become awkward to work with using on-hand database management tools.
- Difficulties include capture, storage, search, sharing, analytics, and visualizing.
- Benefits of working with larger and larger datasets allowing analysts to "spot business trends, prevent diseases, combat crime."
- Scientists regularly encounter this problem in meteorology, genomics, connectomics, complex physics simulations, biological and environmental research, Internet search, finance and business informatics.
- Ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing) "software logs, cameras, microphones, RFID readers, wireless sensor networks and so on."
- One current feature of big data is the difficulty working with it using relational databases and desktop statistics/visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers."
BIG DATA: THE EXTENDED DATA WAREHOUSE

Workload agility & workload complexity

most of the data is structured

Analytic RDBMS appliance or platform

Data velocity & volume

Universal RDBMS

Non-Relational System

most of the data is multi-structured

Data variety & analytic complexity

Source 3: Colin White, BI Research, Big Data Use Cases Matching Technology to Business Requirements
DATA STORAGE & ALGORITHMS ARE MOVING CLOSER TOGETHER

REQUIRES ALL DATA AND ALL ALGORITHMS TOGETHER IN ONE PLATFORM
Sybase IQ In-database Analytics

Simply Smarter

Data To Logic = SLOW + CLUMSY

Logic / filtering applied in Analytic Workbench Servers

Results

Visualization

Logic / filtering applied in In-Database

Results

Visualization

Figure 1: THE OLD WAY

Figure 2: THE NEW WAY

✓ Data NEVER leaves database until results are materialized
✓ Analytics code/models are SHAREABLE and allow AD-HOC analysis
✓ Analytics code/models are applicable to the LATEST data set
✓ PRIVACY protection is ensured
✓ STANDARDS based access, EXTENSIBILITY of In-database logic
✓ PERFORMANCE and SCALABILITY improvements are visible
✓ Average SQL specialist able to code In-database analytical models
Sybase IQ and Fuzzy Logix
Solving Multiple Paradigms

- Balancing between large volumes of data, throughput and accuracy has always been a challenge.
- The conventional wisdom – pick any two (or just one!)
- Sybase IQ and Fuzzy Logix provide an analytical platform where one can achieve all three objectives simultaneously.
- Traditional constraints of data analysis are non-existent in this platform.

*Powered by column based architecture and in-database analytics*
WHAT IS MAPREDUCE?

• Software framework developed by Google in 2004

• Supports distributed computing on large data sets across clusters/nodes

• MapReduce is based on functional programming’s “MAP()” and “REDUCE()” functions

• Libraries written in: C++, C#, Erlang, Java, OCaml, Perl, Python, PHP, Ruby, F#, R and other programming languages
STEP 1: “MAP”

• Master node receives the input

• Partitions input up into smaller sub-problems

• Distributes the sub-problems to worker nodes.

**NOTE:** A worker node may repeat this process, leading to a multi-level tree structure. The worker node processes the smaller problem, and passes the answer back to its master node.
STEP 2: “REDUCE”

• Master node collects the answers to all the sub-problems

• Combines the answers to form the output
WHAT IS HADOOP?

• Software framework owned by Apache

• Enables applications to work with thousands of nodes and petabytes of data.

• Inspired by MapReduce

• Written in Java

• Yahoo! And Facebook are the most prominent users
COMPLEX HADOOP ARCHITECTURE

- Similar to a simple architecture but with dedicated servers for each node
PREDICTIVE MODELING MARKUP LANGUAGE (PMML)

Open Standard Agile Deployment of Predictive Analytics Unifies Proprietary

- Callable via SQL
- In-database scoring
  - Any predictive model
  - Any data mining vendor
- Ease of use
- Highly scalable

Data → RDBMS → Predictions

Model Building

SAP BusinessObjects
Pervasive
IBM
KNIME
R
SPSS
Sas
KXen
MicroStrategy
Salford Systems
TIBCO
Predictive Model Markup Language

Model Deployment Integration / Execution

Simple Deployment & Execution
- Turn PMML into UDFs
- Deploy PMML files & UDF stubs
- Write SQL against UDFs
ANALYTICS INTEGRATED INTO BUSINESS PROCESS PROVIDES REAL-TIME INSIGHTS

NEW ANALYTIC SERVERS RUN PREDICTIVE MODELS AGAINST LIVE DATA IN REAL-TIME PROVIDING MORE ACCURATE, TIMELY INSIGHTS
# TYPES OF IN-DATABASE ANALYTICS

<table>
<thead>
<tr>
<th>Mathematical and Statistical</th>
<th>Time Series</th>
<th>Probability Distributions</th>
<th>Data Mining</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Basic math</td>
<td>• ARIMA</td>
<td>• Monte Carlo simulation</td>
<td>• Linear regression</td>
</tr>
<tr>
<td>• Matrix algebra</td>
<td>• GARCH</td>
<td>• Univariate distributions</td>
<td>• Logistic regression</td>
</tr>
<tr>
<td>• Gamma and beta functions</td>
<td>• Forecasting</td>
<td>• Copulas – Correlated multivariate distributions</td>
<td>• Principal component analysis (PCA)</td>
</tr>
<tr>
<td>• Area under curve</td>
<td>• Auto - Correlation</td>
<td>• Clustering – five models available</td>
<td>• Cluster analysis – five models available</td>
</tr>
<tr>
<td>• ANOVA</td>
<td>• Intervention Detection</td>
<td>• Discriminant analysis</td>
<td>• Discriminant analysis</td>
</tr>
<tr>
<td>• Chi-square</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Descriptive statistics</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Mathematical and Statistical
- Basic math
- Matrix algebra
- Gamma and beta functions
- Area under curve
- ANOVA
- Chi-square
- Descriptive statistics

### Time Series
- ARIMA
- GARCH
- Forecasting
- Auto - Correlation
- Intervention Detection

### Probability Distributions
- Monte Carlo simulation
- Univariate distributions
- Copulas – Correlated multivariate distributions

### Data Mining
- Linear regression
- Logistic regression
- Principal component analysis (PCA)
- Cluster analysis – five models available
- Discriminant analysis
CEP - COMPLEX EVENT PROCESSING

Send the data to the queries.
Streaming data analytics.

**Situation Detection**
- Analyze live data for patterns which identify critical situations

*Examples:*
- Out of range sensor readings
- Out of specification utilization
- Unrecognized IP packets

**Automated response**
- Identify critical situations and drive immediate responses

*Examples:*
- Automated back-testing
- Operational problem remediation suggestions
- Intrusion detection and isolation actions

**Stream Transformation**
- Cleanse and Enrich live data streams
- Turn raw data into actionable info

*Examples:*
- Test data cleansing and enrichment
- Scenario-centric interaction analytics
- Best execution metrics
- Bandwidth utilization analysis

**Continuous Intelligence**
- Track operational performance to adjust execution rapidly

*Examples:*
- Situational Risk Intelligence
- Human and machine fault monitoring
- Dynamic resource reallocation management
MARKET LIQUIDITY ANALYSIS

Consolidates order book across execution venues

- Analyze liquidity in a fragmented market
- Identify when, where to trade
- Algo Trading, Smart Order Routing
REAL-TIME TRADE RISK MONITORING

Analyze live data across multiple dimensions

Counterparty Exposure

Trader Limits

Counterparty Limits

Capital Commitments
ADVANCED ANALYTICS PLATFORM

Technology capabilities combine speed + depth of analysis.

Stream Data
- Ultra-low Latency – answers in milliseconds
  (Complex Event Processing)

Ops Data
- Operational Intelligence – answers in seconds

Other Sources
- High-performance Analytics – answers in minutes/hours

Conventional Database

Columnar Store

Database designed for analytics

SEND DATA TO THE QUERIES
BUSINESS ANALYTICS CHALLENGES

A wide variety of solutions to consider – in fact, one too many

How do I unleash multi-core parallelism: MPP or SMP?

How do I deal with ‘Big data’: Column Store, FPGA, Storage sw?

Parallel Databases OR In-memory databases OR Distributed File Systems?

Look at us!

Appliances, you say?

Why not private clouds?

In-database Analytics?

MapReduce?
DIFERENCIA DE ANALÍTICOS Y OLTP

- El Procesamiento de Transacciones En Línea (OLTP) necesita pocos registros por transacción

- Las consultas analíticas procesan miles de registros
- Solo se necesitan unas pocas columnas de una vez
- Las tablas usualmente son 10x más grandes que el OLTP
- Sybase IQ es óptimo para el inteligencia empresarial debido a su estructura de columnas
NEW APPROACHES – IN MEMORY DB

- Disk is Slower
- Some data case be persisted onto disk
- Less moving parts
- Data recovery issues
- RAM is more expensive making Big Databases not appealing for IMDB

Source
http://www.mcobject.com/in_memory_database
SOLID STATE

Solid State Disks make the process of implementing a complete memory solution simpler by transferring the database tables from the current storage device to a solid state disk and starting up the database. In addition solid state disks provide a host of features that In-Memory Databases cannot match in terms of handling failure.

<table>
<thead>
<tr>
<th>Storage Type</th>
<th>Avg. Read Speed</th>
<th>Avg. Write Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platter Disk</td>
<td>10.0 ms.</td>
<td>7.0 ms.</td>
</tr>
<tr>
<td>DRAM SSD</td>
<td>0.4 ms.</td>
<td>0.4 ms.</td>
</tr>
<tr>
<td>Flash SSD</td>
<td>1.7 ms.</td>
<td></td>
</tr>
</tbody>
</table>

Imdb skips all logging and paging activities... No LRU and MRU or anything like that... Imdb is much more efficient or streamlined than SSD.
MASSIVELY PARALLEL PROCESSING (MPP)

• All resources available for a query
• shared-nothing MPP architecture provides every segment with a dedicated, independent high-bandwidth channel to its disk.
• Fast loading
## COLUMN-BASED ARCHITECTURE

### Conventional Database

- Data is stored horizontally
- Querying without indexes and views is extremely I/O intensive
- Building indexes and views is a huge time and resource drain
- Database footprint must be dramatically expanded to make the environment efficient for querying

### SYBASE IQ

- Data is stored vertically – Each column is stored separately
- The data is the index
- Retrieve only columns used in the query – Reduce System I/O dramatically
- Allocate a thread for each column individually – *Process the query in parallel*
Speed from a column-based architecture
A Basic Query into a Growing Customer Database

An online shopping service has millions of subscribers. Valentine’s Day is coming. How many male customers in California have ever purchased jewelry?

Row-based Database

<table>
<thead>
<tr>
<th>Gender</th>
<th>State</th>
<th>Jewelry Purchases</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>NY</td>
<td>Y</td>
</tr>
<tr>
<td>M</td>
<td>CA</td>
<td>N</td>
</tr>
<tr>
<td>F</td>
<td>CT</td>
<td>N</td>
</tr>
<tr>
<td>M</td>
<td>MA</td>
<td>Y</td>
</tr>
<tr>
<td>M</td>
<td>CA</td>
<td>Y</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

800 Bytes x 10M = 500,000 I/Os

16K Page

- Process large amounts of unused data
- Often requires full table scan

Column-based Sybase IQ

<table>
<thead>
<tr>
<th>Gender</th>
<th>Jewelry Purchases</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>Y</td>
<td>CA</td>
</tr>
<tr>
<td>M</td>
<td>N</td>
<td>CA</td>
</tr>
<tr>
<td>F</td>
<td>Y</td>
<td>NY</td>
</tr>
<tr>
<td>M</td>
<td>Y</td>
<td>CA</td>
</tr>
</tbody>
</table>

10M Bits x 3 Column/8 Bits = 234 I/Os

16K Page

= 2
# TYPES OF SOLUTIONS

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Vendors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional Purpose Built Data Warehouse Solution</td>
<td>Built for DW and Analytics not OLTP</td>
<td>Teradata, Redbrick, Sybase IQ</td>
</tr>
<tr>
<td>Retrofitted OLAP</td>
<td>Built for OLTP and enhanced to support DW and Analytics</td>
<td>Oracle, DB2, MSSQL, Sybase ASE</td>
</tr>
<tr>
<td>Data Warehouse Appliance</td>
<td>HW (disk, memory CPU), OS, and DB all in one</td>
<td>Netezza, Data Allegro</td>
</tr>
<tr>
<td>Hybrid Appliance</td>
<td>OLTP RDBMS built on an appliance with in memory capabilities</td>
<td>Exadata, DB2 Balanced Warehouse *HP Neoview (*abandoned)</td>
</tr>
<tr>
<td>In Memory DW</td>
<td>All data in memory</td>
<td>SAP HANA</td>
</tr>
<tr>
<td>Column Based DW</td>
<td>Data stored/accessed in columns not rows</td>
<td>Sybase IQ</td>
</tr>
<tr>
<td>Column Based Appliance DW</td>
<td>Data stored/accessed in columns not rows sits on a DW appliance</td>
<td>Vertica</td>
</tr>
<tr>
<td>Data Warehouse as a Service</td>
<td>All infrastructure is offsite or cloud based</td>
<td>Kognitio, Vertustream</td>
</tr>
</tbody>
</table>
# HOW SOLUTIONS STACK UP

<table>
<thead>
<tr>
<th>Type</th>
<th>Vendors</th>
<th>Pro/Con</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional Purpose Built Data Warehouse Solution</td>
<td>Teradata, Redbrick, Sybase IQ</td>
<td>Pro – Built for the job nothing else</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Con – Expensive to buy and maintain</td>
</tr>
<tr>
<td>Retrofitted OLAP</td>
<td>Oracle, DB2, MSSQL, Sybase ASE</td>
<td>Pro – Existing apps and skills</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Con – Expensive to maintain</td>
</tr>
<tr>
<td>Data Warehouse Appliance</td>
<td>Netezza, Data Allegro</td>
<td>Pro – All in one, easy to manage</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Con – Married to HW platform, forklift Upgrade</td>
</tr>
<tr>
<td>Hybrid Appliance</td>
<td>Exadata, DB2 Balanced Warehouse *HP Neoview (*abandoned)</td>
<td>Pro – All in one easy for exising apps</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Con – Expensive, one platform</td>
</tr>
<tr>
<td>In Memory DW</td>
<td>SAP HANA</td>
<td>Pro – Memory is faster that disk period</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Con - 5TB off memory? + SW</td>
</tr>
<tr>
<td>Column Based DW</td>
<td>Sybase IQ</td>
<td>Pro – faster than row based, less HW</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Con – Not good at OLTP</td>
</tr>
<tr>
<td>Column Based Appliance DW</td>
<td>Vertica</td>
<td>Pro – faster than row based, all in one</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Con – Immature, may not work</td>
</tr>
<tr>
<td>Data Warehouse as a Service</td>
<td>Kognitio, Virtustream</td>
<td>Pro – No infrastructure to maintain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Con – Your biggest asset is off site</td>
</tr>
</tbody>
</table>
SYBASE IQ – A DBMS FOR NEW ANALYTICS

TECHNOLOGY LEADERSHIP

- State of the art query processor
  - Column store, workload, resource aware

- Versatile column store technology
  - Leading compression, data fetch, schema features

- Data type friendly engine
  - Structured, semi-structured, unstructured data

- Massively scalable grid architecture
  - Highly linear scalability for concurrent workloads
SYBASE IQ 15.3 PlexQ®

MPP – Shared Everything with Distributed Query Processing (DQP)

Characteristics
- Compute and storage resources are decoupled
- Single query can span multiple nodes, disks
- Dynamic, versatile resource allocation
  – Fully shared or private

Advantages
- Economical query scale out across low cost HW units
- Independent scale out of compute and storage
- Simple, non-partitioned data placement
- Scales across spectrum: batch, ad-hoc, concurrent

Addresses: big data, complex queries, analytics for all
Scenario II: concurrent queries mutually exclusive resources
- **Query 1**: Finance queries across revenue, margin data
- **Query 2**: Marketing queries across campaign response data
HADOOP INTEGRATION - SYBASE IQ

Hadoop analysis with relational data and other content.
Native MapReduce API through a SQL interface

Data Integration
Hadoop files loaded via ETL

Data Federation
Hadoop files exposed through DB table

Query Federation
Hadoop queries run via UDF calls
MAPREDUCE - SYBASE IQ

MAPREDUCE METHOD IN SYBASE IQ - UDF-TPF EXECUTION

SELECT new_quotes.* FROM missing_quotes(
    TABLE ( SELECT * FROM iq_quotes_table quotes WHERE ticker IN ('SY', 'TDC', 'ORCL') )
    OVER (PARTITION BY quotes.ticker
        ORDER BY quotes.trade_time)
) AS new_quotes, ...
PMML PLUG-IN FOR SYBASE IQ

Open Standard Agile Deployment of Predictive Analytics

- In-database scoring
  - Any predictive model
  - Any data mining vendor
- Ease of use
- Highly scalable

Model Building
- SAP BusinessObjects
- Periasive
- IBM
- KNIME
- R
- SPSS
- SAS
- KXen
- MicroStrategy
- Salford Systems
- TIBCO

Model Deployment Integration/Execution
- Simple Deployment & Execution
  - Turn PMML into UDFs
  - Deploy PMML files & UDF stubs
  - Write SQL against UDFs

Data → Universal PMML Plug-in → Predictions
ANALYSIS OF BIG DATA – SYBASE IQ

ANALYZE TEXTUAL DATA IN COMBINATION WITH STRUCTURED RELATIONAL DATA

Sybase IQ offers a central analytics grid to analyze all types of Big Data to gain insights across business processes.

Use a variety of techniques simultaneously:
• Text analytics
• Rich media analytics
• MapReduce
• Hadoop

To get clear, easy to digest actionable insight.
Rapid decisions feed ROI. Insight is life blood. Nielsen must provide timely, accurate insights to maintain its position as the guiding voice to the $65 billion TV advertising industry.

Nielsen relies on Sybase IQ's ability to answer even the most complex queries 10 to 100 times faster than traditional data warehouse technologies.

Analyzing large volumes of data (20 terabytes and growing fast) is key to Nielsen’s clients getting the intelligence they need to make the most effective and profitable business decisions and ensure viability.

"I would venture to say Sybase IQ is the fastest Analytics Server in the industry."

Craig Silver, senior database architect of the data technology group at Nielsen
RUN PREDICTIVE MODELS IN-DB

Dramatically accelerate performance, helping meet response time thresholds, and reduce time to implement predictive models

Data mining algorithms:

C++ & Java APIs:
• In-kernel or out-of-process

Types of functions:
• Scalar functions (1-1)
• Scalar aggregates (1-N)
• OLAP aggregates (1-N)
• LOBs
• Table UDFs (1-M rows)
• Table Parameterized UDFs (N-M rows)
SYBASE IQ

COMPREHENSIVE PLATFORM FOR BUSINESS ANALYTICS

Eco system

- Optimal BI, EIM, Model, Replicate
- Dev and admin tools
- Predictive Analytics
- Packaged ILM apps

App Services

- Comprehensive ANSI SQL w/ OLAP
- Built-in Full Text Search
- InDB Analytics w/ MapReduce + simulator
- Web 2.0 APIs
- BigData OpnSrc APIs

DBMS

- Most mature column store
- Comprehensive lifecycle tiering
- MPP queries + Virtual Marts + User scaling
- High Speed loads
- Structured + Unstructured Store
CUSTOMER CASE STUDY

A joint-venture between Belgacom, Swisscom and MTN.

CHALLENGES

• Act proactively on possible future events that would threaten their customers’ quality of service
• Immediate visibility into service problems
• Flexible real-time aggregation of data

RESULTS

• Quality guarantee thanks to real-time, proactive response to emergency signals in periods of peak traffic
• Real-time voice network monitoring and analytics for high QoS
• Comprehensive network monitoring and visibility
• Rapid, flexible analysis of real-time data on demand

“...The gain in storage space generated by Sybase CEP’s data aggregation saves us from having to make substantial investments in additional servers, CPU, power, a more extensive database infrastructure, making major investments that weren’t very future-proof would have meant going through the same process again a few years later.”

Laurent Missa
OSS Manager at BICS
Sound ROI hinges on targeting the right customer

Sybase IQ enables advanced predictive models providing customer analytics for large scale marketing campaigns

The solution crunches through customer behavior, purchasing history, and many other relevant metrics

Sybase IQ lets Playphone gain significant competitive advantage by providing new capabilities in:
  - customer targeting
  - operational efficiency
  - fraud detection

“Sybase IQ is a brilliant analytics engine. I don’t know that the business would still be here today without our Sybase solutions.”

Simon Rose, Director Of Infrastructure, PLAYPHONE
Increased revenue hinges on understanding each customer

Sybase IQ and KXEN deliver the next best action for each customer via product recommendations

Increased customer conversion rate by more than 300%

Grew revenue per transaction over 150%

“HMV sells tens of thousands of items via its recommendation system, Sybase IQ and KXEN’s InfiniteInsight™ offer the right functionality at the right cost”

Hideki itchikawa, information system development manager, hmv
AOK HESSEN

INSIGHTS FROM BIG DATA ANALYSIS

• AOK Hessen, must search for patterns across its extensive existing data sets quickly and simultaneously
  • medical treatment
  • prescriptions
  • insurance benefit claims

• Sybase IQ significantly enhanced AOK Hessen’s ability to handle complex business predictions involving multi-dimensional analysis of many input variables

“The divisions currently using the tool run a significantly greater number of analyses than ever before. They keep discovering new ways of drilling down into data while working with the software.”

Michael Shimmelpfennig, Service Manager, IT-Business Department, aok hESSEN
U.S. Internal Revenue Service reaps 995% ROI from 350+ TB Sybase IQ-powered system helping with fraud detection. Evaluate 24,000 variables in analysis.

"Sybase IQ is our secret weapon."

The largest airline in the world - American Airlines IQ runs its ticket fraud system on Sybase IQ – saving them $5MM annually.

"Sybase IQ was consistently 10 times faster."

SEB Bank runs a 2.2 billion row database on Sybase IQ for over 9 years without interruption.

comScore Networks runs one of the world’s largest data warehouse (1.7 trillion rows) on Sybase IQ and exceeds 40% compression rate 300TB of data in 150TB storage).

"We needed a database that could scale very quickly, but we also needed one that was cost-effective."

Sybase IQ replaced Teradata at Atos Worldline saving 60% in costs, improving query speed by 80%, and improved concurrent scalability.
ADDITIONAL READING

• McKinsey Report
  “Big data: The next frontier for innovation, competition, and productivity”
  http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

• TDWI Best Practices Report
  “Big Data Analytics”

• The Economist
  Lots of articles
  Example: “Building with big data”
  http://www.economist.com/node/18741392