Real-Time and Big Data Challenge Data Management Best Practices

Philip Russom
TDWI Research Director for Data Management
May 1, 2012
Speakers

Philip Russom
TDWI Research Director, Data Management

Marie Goodell
Sr. Director for Solution Marketing, SAP
Today’s Agenda

• The Mega Trends
  – Competing trends
  – Real-time versus big data, complicated by interoperability
  – May lead to compromises in data management best practices

• Data management strategies without compromise
  – Moore’s law, selecting the right data management discipline, new data management architectures and options, integrated tool platforms, using more real-time data management functions

• Recommendations
The Mega Trends

These are today the most influential trends in data-driven IT disciplines, e.g.:
Business intelligence, data warehousing, data management, analytics...

• Real-Time Operation
• Scaling Up/Out to Big Data
• Interoperability among IT Systems
• Do More With Less
Big Data versus Real-Time Operation
With Interoperability Caught in the Middle

“Do More With Less” – All you need is regulation rope & matching uniforms!
Big Data Volume: TBs today, PBs soon

What’s the approximate total data volume that your organization manages ONLY for analytics, both today and in three years?

<table>
<thead>
<tr>
<th>Data Volume Range</th>
<th>Today</th>
<th>In three years</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1 TB</td>
<td>8%</td>
<td>2%</td>
</tr>
<tr>
<td>1–3 TB</td>
<td>14%</td>
<td>4%</td>
</tr>
<tr>
<td>3–10 TB</td>
<td>22%</td>
<td>12%</td>
</tr>
<tr>
<td>10TB–100TB</td>
<td>37%</td>
<td>31%</td>
</tr>
<tr>
<td>100TB–500TB</td>
<td>6%</td>
<td>21%</td>
</tr>
<tr>
<td>&gt;500 TB</td>
<td>5%</td>
<td>20%</td>
</tr>
<tr>
<td>Don’t know</td>
<td>8%</td>
<td>10%</td>
</tr>
</tbody>
</table>

- Users conduct analytics with ever-larger datasets.
- Small analytic datasets will become less common, as they grow into large ones.
- A third of surveyed organizations (37%) have broken 10Tb barrier.
- Very large analytic datasets will become much more common.
- Soon we’ll measure big data in petabytes, not terabytes.

SOURCE: 2011 TDWI report Big Data Analytics
Big Data isn’t the problem it used to be.

In your organization is big data considered mostly a problem or mostly an opportunity?

- **Problem** – because it's hard to manage from a technical viewpoint
- **Opportunity** – because it yields detailed analytics for business advantage

Only 30% of survey respondents consider big data a problem.
- *Oddly enough, big data was a serious problem just a few years ago.*
- *Storage and CPUs developed greater capacity, speed, intelligence*
  - They also fell in price.
- *New database mgt systems arrived, designed for big data analytics.*
  - Data warehouse appliances and analytic DBMSs are scalable and relatively affordable.

The vast majority (70%) considers big data an opportunity.
- *The recent economic recession forced deep changes in most businesses.*
- *Big data analytics reveals change’s root cause, so you can stop it or leverage it.*

Source TDWI. Survey of 325 respondents, June 2011

**SOURCE:** 2011 TDWI report Big Data Analytics
Why is Real-Time a Tech Challenge? Because it takes time to...

- To move data
  - *Connect to a source system*
    - Typically DI tool connecting to operational app
  - *Runs a query or otherwise identifies data*
  - *Extracts data into a network package*

- To prepare data properly
  - *Join data from multiple sources*
  - *Data quality functions: validate, standardize, dedupe, enhance*
  - *Goal is to deliver clean, compliant, complete, contextual, auditable data*

- To load data
  - *Connect to source, run updates and inserts*

- To present data
  - *Refresh a report/dashboard, alert appropriate user*
Data Integration is caught between A Rock and a Hard Place

- The two requirements run the risk of cancelling each other out.
- How can you satisfy both requirements, without sacrifice?

**ROCK:** Business Needs Time-Sensitive Data Delivered Quickly

**HARD PLACE:** Quality Data from the Best Sources Presented Well

For DI solution, it takes time to:
- Find data
- Move data
- Prep data
- Load data
- Present data…
A COUPLE OF EXCEPTIONS
Sometimes, Junk Data is all the Business Needs

• Fuzzy Metrics
  – Many metrics and KPIs are approximations
  – If a metric is intended to be “fuzzy,” then there’s no point in improving its accuracy
  – Likewise, if the metric is purely tactical information, and no strategic decisions are based on it

• Preliminary Analytic Results
  – Much of the discovery work business analysts do is not a deliverable & not intended for broad consumption
  – However, after an analytic epiphany, analytic findings must be operationalized with quality data, reports, etc.
Options for Real-Time Data Management

*For users, there are many high-performance options available today.*

- **Real-Time Data Integration** – Next Gen Hot Growth Features:
  - *RT DI, RT DQ, complex event processing (CEP), RT alerts*
- **Real-Time Data Quality**
  - *Standardize, validate, enhance data before it hits database*
- **Real-Time Master Data Management** – Next Gen Hot Growth:
  - *Instantiate 360-degree customer view or publish ref data in RT*
- **Data Replication** – According to Next Gen DI Survey:
  - *46% of users do replication; second only to ETL!*
- **Data Federation** – According to Next Gen DI Survey:
  - *33% of users surveyed are now using federation. It’s arrived.*
- **Data Services over an Enterprise Service Bus**:
  - *Reach operational application data in RT; 39% w/data services*
- **In-Memory Databases**
  - *With multi-Tb memory, you can now put Big Data in memory*
- **In-Database Analytics**
  - *Faster to analyze big data where it’s stored, instead of moving it*
Why Care About Real-Time Data Management?

• It’s a foundation for time-sensitive business practices:
  – Operational business intelligence
  – Just-in-time inventory
  – Facility monitoring
  – Self-service information portals
  – eCommerce recommendations
  – Price optimization
  – Production yield & workforce mgt in manufacturing

• Real-Time Analytics is coming
  – Reporting embraced RT. Just look at Operational BI.
  – Analytics, too, will soon take that trip. And with Big Data.
Use Cases for Data Federation

• Data Warehousing
  – *Prototype before you persist*
  – *Augment DW with current data*
  – *Federate DWs and marts*
  – *Migrate data warehouses*
  – *Augment ETL*

• Other
  – “SWAT” applications
  – Operational reporting
  – Data services layer in SOA
  – Data discovery and modeling

Source: TDWI
“Do I really need DI or DW for real-time data?”

- Yes. Include data integration in the process, because it:
  - Transforms complex data into useful and reusable forms
  - Improves data’s quality, metadata, master data
  - Comply with policies for data access, use, security, privacy

- Yes. Include a data warehouse in the process because it:
  - Provides a complete historic context to complement real-time data
  - Maintains an audit trail
  - Records the evolving state of real-time data

- Omit DI and DW at your peril:
  - It’s possible to get current data by going around DI and DW.
  - But the data won’t be clean, compliant, complete, contextual, auditable…
In-Memory Databases as Accelerators

• A use case in OLAP
  – *Put OLAP cube or table(s) in server memory*
  – *Update memory cache as data evolves*
  – *Users can refresh reports or dashboards from memory cache, in real time and on demand*

• A use case in predictive analytics
  – *Put analytic models and data in memory*
  – *Rescore models quickly, so in real time you can*
    • Track changes in consumer behavior
    • Test for fraud and risk

• Related Trends
  – *32-bit to 64-bit upgrades*
    • Many are driven by need for 64-bit’s massive addressable memory space
  – *BI & DW Appliances*
    • Many are designed to be an in-memory DB, sometimes called an accelerator
  – *Flash drives and solid-state drives are coming for BI/DW servers*
    • Persistent storage (like disks), but with high speed (like memory)
In-Database Analytics Simplifies Data Warehouse Architecture

OLD WAY
Dump, Score, and Load

Dump or ETL Data Out
Rescore Analytic Models

Scores

Analytic Data

DW or Other Database

Load Scores for Use

NEW WAY
In-Database Analytics

Dump or ETL Data Out
Rescore Analytic Models

Scores

Analytic Data

DW or Other Database
TDWI SURVEY SEZ:
Users would prefer an Integrated DM Platform

- Users with multiple DI tools from multiple vendors will drop from 44% to 25%.
- 9% report using integrated suite of tools one today, yet 42% would prefer one.
- Integrated data management platforms are the future.
  - *Greater reuse of data mgt artifacts, since they’re shared across multiple tools*
  - *Enables users to build solutions incorporating multiple data mgt functions*
    - Easier to squeeze data quality functions into more solutions
    - Easy access to multiple real-time data mgt functions

Which of the following best describes your organization’s portfolio of DI tools today? For your organization’s next generation DI implementation, how would you prefer that the DI portfolio be?

<table>
<thead>
<tr>
<th>TODAY</th>
<th>WOULD PREFER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using multiple DI tools from multiple vendors</td>
<td>44%</td>
</tr>
<tr>
<td>Using just one DI tool</td>
<td>22%</td>
</tr>
<tr>
<td>Mostly hand coded without much use of vendor DI tools</td>
<td>18%</td>
</tr>
<tr>
<td>Using a DI tool that’s part of an integrated suite of data management tools from one vendor</td>
<td>9%</td>
</tr>
<tr>
<td>Using multiple DI tools from one vendor</td>
<td>3%</td>
</tr>
<tr>
<td>Other</td>
<td>4%</td>
</tr>
</tbody>
</table>

SOURCE: 2011 TDWI report *Next Generation Data Integration*
TDWI SURVEY SEZ:
Users will employ more Real-Time DI functions

- Users employ 40% of their DI tool’s functions, on average.
- This will jump to 65% of DI tool functions.
- Real-time DI functions will see the most growth.

“What approximate percentage of your primary DI tool's functions are you using?”

SOURCE: 2011 TDWI report Next Generation Data Integration
Miscellaneous Observations & Caveats

- Not all data needs to travel in real time or be analyzed
  - *Work with biz; provide only what they need for operational decisions*

- If you over-prepare analytic data, you can lose nuggets
  - *Some adv’d analytic apps depend on raw, detailed source data, as in Big Data*
  - *Un-processed source takes much more time to process than aggregated data*

- You really do need to mix RT operational data with DW data
  - *RT Op data represents the moment, DW data the historic context*
  - *Without both, you don’t have the complete picture*
Recommendations

• Be aware of the Mega Trends for data management
  – Plan accordingly, esp for real time, big data, interoperability

• Embrace Big Data
  – It’s not the problem it used to be

• Expand your use of data management tools & functions
  – Embrace real-time interfaces, services, buses

• Rely on key technology enablers
  – Real-time DI, DQ, MDM, events
  – Federation, replication, in-memory DBs

• Give the business what it needs
  – Faster, broader data for operational decisions
  – In future, more real-time analytics, too
Managing Big Data to Gain Insight Never Seen Before

Marie Goodell
How To Manage Big Data Effectively?  
...With a real-time data platform

- Eliminate Unnecessary Data Management Costs
- Accelerate Performance and Improve Business Processes
- Unlock New Business Opportunities

SAP Real-time Data Platform

Volume

Velocity

Variety

VOLUME

VALUE

VELOCITY

VARIETY
SAP Real-Time Data Platform
A flexible deployment environment for managing big data
SAP Data Management Portfolio
Flexible Storage & Processing Options for Your Business Needs

Embedded

Transactions

Open EDW

Continuous Intelligence

Real-time Business

Transactional Database, Best TCO

Open Analytical Database, Best TCO

Insight into Streaming Data

Next-Gen Platform for Real-time Business

Sybase SQL Anywhere

Sybase ASE

Sybase IQ

Sybase ESP

SAP HANA

• Leading mobile and embedded database
• True zero administration; Best store and forward synchronization
• Optimized for extended enterprise, ISVs for cloud and device based apps

• Database for extreme transaction applications with best TCO
• High reliability and low maintenance
• Optimized for Purpose Built Apps, Business All-in-One, SAP ERP

• Smart "Big Data" Store with best TCO
• High performance, low price / performance
• Optimized for: Open Enterprise Data Warehouse

• #1 Complex Event Processing (CEP) platform
• Faster, simpler way to analyze and act on event streams;
• Optimized to support real-time data warehouse; sense & response application

• Transactional, analytical and application logic processing in ONE DBMS
• Best real-time data mgmt – eliminate redundant layers, latency, modeling (OLAP on OLTP model)
• Optimized for BW on HANA, HANA as data mart, HANA apps
SAP HANA Data Platform
Ideal for real-time business

1. Revolutionary in-memory platform
   • Real-time analytics on detailed data on the fly
   • In-memory calculations
   • Real-time replication to eliminate data latency
   • No aggregates, tuning of data for performance

2. Empowers you to interrogate data
   • Wizard-driven data modeling for business
   • Fast & easy creation of ad-hoc views
   • Optimized for SAP BusinessObjects BI
   • Open platform for other clients

3. Powerful predictive analytics
   • Embedded data mining algorithms for predictive analytics
   • Bring decision support capabilities to the business users through simplified experience and pre-built scenarios

4. High data quality
   • Real-time replication
   • Faster integration with optimized ETL
   • Tightly connected with EIM data quality capabilities
SAP EIM Portfolio for Big Data
Ensure complete and accurate information

Collaborative IT / business
Design your information architecture

Available
Replicate data rapidly

Complete
Integrate & process data from a variety of stores

Accurate
Cleanse & enhance data

Single view
Ensure master data quality

Trustworthy
Profile and measure data quality

SAP Data Services – Data Quality Mgmt
- Assess, correct, complete data
- Validate & enrich data
- Quality coverage for 230 countries
- Geo-location / address cleansing

SAP Enterprise Master Data Mgmt
- Manage master data created in business apps
- Manage master data across distributed systems
- Product, material, supplier, customer
- Global data sync

Sap PowerDesigner
- Data modeling
- Information modeling
- Enterprise architecture modeling
- Common metadata repository

Sybase Replication Server
- Move & synchronize data
- Disaster recovery
- High availability
- Load balancing
- Sybase, Oracle, IBM, Microsoft

Sybase PowerDesigner
- Rapid ETL
- Read & load Hadoop data
- Process text data to unlock meaning from unstructured documents

SAP Information Steward
- Graphical UI for business users
- Data quality scorecards
- Data lineage
- Change management impact analysis
- Common metadata repository

Accurate
Cleanse & enhance data

Available
Replicate data rapidly

Complete
Integrate & process data from a variety of stores

Trustworthy
Profile and measure data quality
SAP Data Services – Hadoop Connector
Connect unstructured & structured data for greater insight

**Map-Reduce, is a parallel processing paradigm where code is sent to data for instant processing**

1. **Collect & Store**
   - Files are stored in their native format incurring no transformational costs.
   - Built in fault-tolerance.
   - Commodity hardware and software solution makes Hadoop scale cost effectively.

2. **Analyze & Process**
   - Prepare the data to enable the class of problems to be solved.
   - Problems like searching, counting, pattern detection lend themselves well to Map Reduce paradigm.

3. **Integrate & Consume**
   - Flexibility and change management; security and access control
   - Administrative and operational skill set for business continuity
   - Enterprise support and Integration into enterprise infrastructure

Log Files
Multi-structured Data Sources
Media and Other Files

SAP HANA
Data Warehouses
Structured Data
Disk

Enterprise Portals
Mobile
Dashboard/Report
On Demand Services
Steven Paul Jobs (born February 24, 1955) is an American business magnate and inventor. He is well known for being the co-founder and chief executive officer of Apple. Jobs also previously served as chief executive of Pixar Animation Studios; he became a member of the board of The Walt Disney Company in 2006, following the acquisition of Pixar by Disney.
SAP BusinessObjects Business Intelligence platform
Access, analyze and share information from big data

**Empower all people, enable all workflows**

**Common user experience for all front-ends**
- Crystal Reports
- Web Intelligence
- Explorer
- Dashboards
- Predictive Analysis

**High performance, feature rich and secured access**

**Best access method for each specific data source**
- Universe Access
- Direct Access

**The new information design tool is your point of entry to business intelligence solutions**

**All data sources**
- SAP BW
- Sybase
- SAP HANA
- Any Relational Database
- Files
- Web Service
1. Understand the business and identify issues
2. Load the SAP and non-SAP data into SAP HANA or Sybase IQ using SAP Data Services

**Step 1: Data Loading**

1. Understand the business and identify issues
2. Load the SAP and non-SAP data into SAP HANA or Sybase IQ using SAP Data Services.

**Step 2: Data Preparation**

1. Data examination
2. Data visualization with SAP BusinessObjects Predictive Analysis
3. Sample, filter, merge, append, apply formulas

**Step 3: Data Processing**

1. Define the model - clustering, classification with association, time series etc. (SAP BusinessObjects Predictive Analysis)
2. Run the model in SAP HANA or Sybase IQ

**Step 4: Data Visualization**

1. Visualize the model for better understanding with SAP BusinessObjects Predictive Analysis.
2. Store the model and result back to SAP HANA or Sybase IQ.
Flexible Real-time Data Platform
Based on business needs, speed, and cost

SAP HANA for Real-time Business
- Built for extreme transactional, analytical and application logic processing in ONE database
- Best real-time data management - eliminate redundant layers, latency and modeling (OLAP on OLTP model)

Sybase ESP for Streaming Events
- CEP for monitoring & processing streaming events in real time
- Best for immediate action

Sybase ASE for Extreme Transactional Database
- Built for transaction applications
- Simplest maintenance
- Best TCO for SAP application + database

Apache Hadoop for Distributed file system based storage and processing:
- Open source
- Emerging storage and processing model for unstructured data

Sybase IQ for Open Analytical Database
- Built for data warehousing
- Operate with any technologies (e.g. HW, Hadoop)
- Best cost/performance in market

Sybase DB / DW
- Cost-effective Distributed FS

In-Memory Innovation

Business Analytic

Cost-effective Distributed FS

In-Memory Innovation

Transactional Management
THANK YOU
Questions?
Contact Information

If you have further questions or comments:

Philip Russom, TDWI  
prussom@tdwi.org

Marie Goodell, SAP  
marie.goodell@sap.com