In-Memory Data Fabric: A Modern Approach to Data Warehouse Architecture

Philip Russom
TDWI Research Director for Data Management
June 25, 2014
Sponsor
Speakers

Philip Russom
TDWI Research Director,
Data Management

Tom Traubitz
Senior Director, Product Strategy
SAP
Agenda

• Modern Business Practices
  – Faster and broader data mgt
  – Lean and agile methods
  – Real-time operations
  – Embedded analytics, etc…

• Key Emerging Technologies
  – Data Virtualization
  – In-Memory Data & Processing

• Data Fabric
  – Combination of in-memory functions and virtualized data
  – Special case: Logical Data Warehouse

• Conclusions & Recommendations

PLEASE TWEET @pRussom, @SAP, #TDWI, #DataFabric, #LogicalDW, #Analytics, #RealTime
Modern Business Practices

• Keep pace with accelerating business processes
  – See that a customer is about to churn; immediately reach out to retain him
  – Monitor operations for greater speed, yield, quality
  – Compete by being faster than competitors

• Software automation can accelerate organizations
  – BI, DW & analytics are key to automating speed
  – Businesses are pushing BI/DW closer to real time

• Operationalizing analytics is another biz trend
  – Organizations are integrating and embedding BI/DW & analytics into more operational apps

• But it’s not just real-time speed…
Data-Driven Business Practices

• Run the business by the numbers
  – Requires fresh data, from best sources, delivered fast, to key people
• Complete information
  – Complete customer views, enterprise-scope data, social media, big data…
• Predict the future via analytics
  – This is the next step for many orgs
  – Goes beyond real-time operations
• Operational Excellence
  – Efficiency & accuracy result from practices that are data-driven, real time, lean, agile, based on all data…
Data Management is caught between A Rock and a Hard Place

- 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 Best Sources delivered to the Right People

For DM solution, it takes time to:
- Find data
- Move data
- Prep data
- Load data
- Present data…
For technology, it’s like a tug of war...

...but the two can be reconciled successfully.
Reformation of Data Mgt Practices

• Less a priori data preparation.
  – *Do more ETL, staging & aggregation on the fly, as data’s needed.*

• Less up front modeling and transformation.
  – *Remodel & transform data on the fly, close to real time, as form of late binding.*

• Less bulk data cleansing.
  – *Instead of scrubbing whole databases regularly, cleanse data as it is created, accessed or updated.*

• Less data movement.
  – *Instead, more direct access to operational data (& Hadoop) in real time or close to it.*

• Less redundancy of data stores.
  – *Instead of generating yet another mart or data store, access sources, as needed; join and aggregate on the fly.*

• Free up compute cycles for:
  – *More data, fewer samples.*
  – *For more complete views, better insights.*
ENABLING TECHNOLOGIES for REFORMED DATA MGT

Data Virtualization

• Purely semantic views of data structures
  – No physical data, until view is materialized
• Benefits of data virtualization
  – Doesn’t prep & persist a lot of data on the off chance a user or app might need it
  – Collects fresh data, as needed, instead of hoarding stale data
• Various processing available
  – Some views are read only
  – Others can write data and perform data processing functions (or call them)
    • E.g., views that represent joins or aggregates
• Virtualization intersects with real time
  – Most views (but not all of them) execute in real time (or close) when materializing data
ENABLING TECHNOLOGIES for REFORMED DATA MGT

In-Memory Data Functions

- Data mgt & processing in server memory
  - *Rarely a DBMS in memory*
  - *Usually a data subset in memory*
- Benefits of in-memory data
  - *Eliminates disk IO, which is traditional bottleneck for data mgt*
  - *Provides high performance for many data-driven applications, including data virtualization*
- Various processing available
  - *Simple table cached in memory*
    - E.g., table of metrics/KPIs for dashboards
  - *Multidimensional data*
    - E.g., cube of sales data for intraday analysis
  - *Analytic models and scores*
    - E.g., rescored intraday to spot/report likely churn
ENABLING TECHNOLOGIES
for REFORMED DATA MGT

In-Memory Data Fabric

• Data Fabric is a unified view (or collection of views) of data in multiple systems across an enterprise
  – Plus a simplified (yet diverse & performing) collection of interfaces into such sources and targets
• The point of a data fabric is to provide:
  – A fairly comprehensive big picture of enterprise data
  – A single layer through which data can be accessed, thereby reducing data redundancy, movement, processing
  – A simplified view & mechanism that enables more user types
• In-Memory Data Fabric (IMDF) is combination of things:
  – The data fabric, in-memory data functions, and data virtualization discussed earlier, integrated w/usual apps, databases, & data mgt tools
• Benefits of IMDF
  – A high-performance form of a data fabric, due to in-memory data functions, parallel processing, direct interfaces, optimization, etc.
  – Real-time speed for time-sensitive biz practices, lean data mgt, scalability, embedding analytics in apps, operationalization, etc.
Data Fabric – Example 1
Data Fabric – Example 2
Use Cases Enabled by In-Memory Data Fabric

Real-Time Business Practices

• Real-Time technologies are a foundation for time-sensitive business practices:
  – Operational business intelligence
  – Just-in-time inventory
  – Facility monitoring
  – Self-service information portals
  – eCommerce recommendations
  – Production yield & workforce mgt in manufacturing

• Real-Time Reporting is common
  – Real-Time Analytics is coming on strong
Use Cases Enabled by In-Memory Data Fabric

Near-Time Analytics

• Retail Optimization
  – *Commodity pricing; short-term sales and bundles*
  – *Inventory shuffle, just-in-time inventory*

• Loans, credit cards, insurance
  – *Immediate review of applications submitted online*

• Visibility into Operational Processes
  – *Yield, quality, efficiency, SLA*
Use Cases Enabled by In-Memory Data Fabric

Visibility and Awareness

- **Visibility** = Know and act on the knowledge quickly:
  - *Know and correct SLA or performance problems*
  - *Spot and stop fraud or security breaches*
  - *Feel confident, knowing that “all systems are go”*

- **Situational Awareness** = See & react accordingly:
  - *See a cluster of street crimes*
    - Deploy squad cars as a deterrent.
  - *See a drop in unit production on manufacturing floor*
    - Bring in more workers and turn on more machinery.
  - *See a product recurring in abandoned shopping carts*
    - Run a promotion to close more sales of that product.
  - *See a social media sentiment or pattern*
    - Direct it or correct it as it evolves.
Recommendations

• In-memory data functions & data virtualization are more viable than ever – so use them!
  – *New level of maturity for speed, reliability, functionality, interoperability*
• Put in-memory functions & data virtualization together in a data fabric
  – *Use in-memory functions for speed and as a point of integration*
  – *Use data virtualization for agile dataset design in development and integration on the fly in deployment*
• Design a data fabric as the primary, central representation of enterprise data
• A data fabric should be an access layer
  – *Not just an academic representation of data*
• Some organizations may need multiple overlapping data fabrics
• Note that a data fabric is seldom 100% virtual, in-memory, real time, or enterprise scope
  – *Based on your organizational needs, selectively decide which data is represented in which fabric, plus which data needs to be virtual or real time*
• For greatest success with a data fabric, infuse it with ample data virtualization and in-memory caching and processing.
  – *The result is an In-Memory Data Fabric (IMDF)*
Reinventing the Data Warehouse
What if, the data warehouse was so fast that…

- No advanced data organization required
- No pre-compute aggregates or pre-stage requirements
- Access to the warehouse didn’t need to be limited
- You could store all the data you want
- Data & events could be captured without intermediate databases
And, the data warehouse could…

**INSIGHT**

- Intelligently access data outside itself
- Manage business events in real time
- Come with powerful, easy-to-use modelling tools
- Have powerful tools for cleansing and extracting data
- Handle all types of data including geo-spatial and text
Simplified, Accelerated, Predictive

Traditional: OLTP and OLAP Separate

Transactions

ETL

Multiple Data Sources

Streams

Staging DB

48/hr Old Data

EDW

Old Data

Immediate

Transactions

Streams

OLTP + OLAP in SAP HANA

Current Data

Smart Data Access

SAP HANA

Multiple Data Sources available with Live Access

Lots of separate ETL processes!

10:00 AM

48 Hours

10:00 AM

10:00 AM

10:00 AM
Reinventing the Data Warehouse with an In-Memory Data Fabric

Results at the Speed of Memory

Data Fabric Layer

- In-memory Platform
- Orchestration
- SQL or SAP River
- Business Applications
- Tightly integrated orchestration for management, monitoring, and control

Real-time Events/Machine-generated Data
- Column Storage
- Petabytes of Structured Data
- Op RDBMS
- Other Sources
- Load Source Databases

Stream
- SDA
- MapReduce/Hive
- ETL & Rep for RT sync
World’s Largest Data Warehouse – NEW Guinness World Record

Audited Record: 12.1 Petabytes

Tested Configuration

22x HP ProLiant DL580 G7
- 4x Intel Xeon E7-4870 @ 2.40GHz
- 1TB RAM

20x NetApp Storage Arrays E5460s
- 60/120 x 3TB 7.2Krpm HDD
- 4 x Fibre Channel connections

SAP IQ 16 (20 nodes)

SAP HANA (5 nodes)

BMMsoft Federated EDMT 9 with UCM

Red Hat Enterprise Linux 6.4 X86-84

http://www.guinnessworldrecords.com/world-records/5000/largest-data-warehouse
Real-time insights mean real-time results

100% accuracy in early signal detection

216x faster DNA results from 2 days to 20 minutes

$1.1M increased revenue with 1% increased retention rate

500k Euro working capital reduction with 1 week

50,000 daily sports betting games analyzed in real-time

3.2M reclaimed by identifying fraudulent insurance charges
Isn’t It Time To Reinvent Your EDW Strategy?

SAP In-Memory Data Fabric
A Complete EDW Architecture for All Your Knowledge Workers
Thank you

Contact information:

Tom Traubitz
tom.traubitz@sap.com
Questions?
Contact Information

If you have further questions or comments:

Philip Russom, TDWI
prussom@tdwi.org

Tom Traubitz, SAP
tom.traubitz@sap.com