New Waves in Business Intelligence and BI Best Practice in IBM

Roy Elgar, AsiaPacific Data Warehouse Architect, IBM Software Group
September 17, 2004
Introduction: Roy Elgar: IBM Software Group

Data Warehouse Architect
AsiaPacific Advanced Technical Sales

relgar@nz1.ibm.com
+64 4 576 8152 office
+64 21 488 143 mobile
New waves in Business Intelligence

Everything is derived from views of transactions
Data Warehousing current waves:

- Consolidation
- Real-time (Right-time) Business Intelligence
- Extending the Data Warehouse via Enterprise Information Integration
- Business Performance Management
Traditional Centralized Warehouse

BI Apps

Application

Data Marts

Mart

Data Warehouse

ODS Layer

ETL / Replication

ETL / Replication

ETL / Replication

ODS

Issues:
Multiple Copies
Time Delays
Adds to overall TCO
Difficult to change

Operational systems
Consolidated Centralized Warehouse

BI Apps

Application

Data Marts (LOB Apps)

Atomic Data (normalized)

ODS / Staging Layer

ETL

Operational systems

Logical and or Physical Mart layer
User playpens and multiple databases

- Application
- Application
- Application
- Application
- Application

Owned and Maintained by LOB Application Developers

Owned and Maintained By Core Warehouse DBAs

App Timeline
Add incrementally

EDW

User Playpen
Separate Database
Data Warehousing current waves:

- Consolidation
- Real-time (Right-time) Business Intelligence
- Extending the Data Warehouse via Enterprise Information Integration
- Business Performance Management
Real Time BI Applications

- **Customer Touchpoint Analytics**
  - Call-center online customer scoring
  - Kiosk in-store coupons
  - ATM personalized advertising
  - Web-site.COM personalized offers & coupons

- **Supply Chain continuous forecast optimization**
  - Pipeline monitoring, Just-in-time inventory

- **Fraud detection**
  - ATMs, cell phones, POS shrinkage

- **Airline ticket pricing or cancelled flight rerouting**
- **Online mortgage lending credit approval analysis**
- **Telco/Energy OnDemand network/electric routing**
- **Train, Truck, Ship cargo packing & routing**
Real time BI

Parallel ETL Engines
MQSeries queues
Replication
Web services

DB2 Warehouse
Information Integration

Concurrent User Queries

Personalization, data mining, rules, campaigns
Alerts, triggers, KPIs, Analytics
Consumers
Corporate Dashboards

Concurrent Loading
## Sprint PCS

<table>
<thead>
<tr>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>360 degree customer view</td>
</tr>
<tr>
<td>Churn prediction</td>
</tr>
<tr>
<td>Unified customer contact</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Business Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campaign responses up 66-300%</td>
</tr>
<tr>
<td>Margin per customer up 20%</td>
</tr>
<tr>
<td>Fraud detection &lt; 4 hours</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB2 EEE, SAS,</td>
</tr>
<tr>
<td>512 CPU pSeries</td>
</tr>
<tr>
<td>Load 250M CDRs/day, Fact table 16TB</td>
</tr>
<tr>
<td>1000s concurrent users, 7000+ customer care</td>
</tr>
</tbody>
</table>

### Call Center
- Fraud Detection
- Network Analysis
- Marketing

- DB2 EEE 171TB (including HA mirror)

**Continuous data loading**
Data Warehousing breaking wave:

- Consolidation
- Real-time (Right-time) Business Intelligence
- Extending the Data Warehouse via Enterprise Information Integration
- Business Performance Management
Three basic models for extending the DW with Enterprise Information Integration:

1. Joining real time data to the Data Warehouse

2. Accessing XML and unstructured data

3. Joining Data Marts to the Data Warehouse
Federation: Join real-time data to the warehouse

- Access current customer records from a call centre
- Access current stock levels from a supply chain data mart
- Business activity monitoring – linking events to trends
Federation: Access XML & Unstructured Content

- Access to customer documentation (e.g. letters, media) from a call centre
- Linking photos or documents to analysis of customer claims in insurance
Federation: Joining Marts & Warehouses

- Access to marts developed by different departments for specific data
- Access from mart to warehouse level for detailed data
Best Practice:  
Start small and grow!

- Where to start
  - Pilot project
  - Around the core of an existing Data Warehouse or ODS
  - Accessing relatively small amounts of remote data
  - Minimize query complexity & focus on queries that remote systems were designed to handle anyway

- What to avoid
  - Simultaneous warehouse & federation projects
  - *Ad hoc* user queries
  - Very complex joins over many sources
Data Warehousing breaking wave:

- Consolidation
- Real-time (Right-time) Business Intelligence
- Extending the Data Warehouse via Enterprise Information Integration
- Business Performance Management
Intersecting Strategies

Business Performance Management

- Key Performance Indicators
- SCM, ERP, CRM
- Balanced Scorecard
- Business Activity Monitoring
- Zero Latency Enterprise
- Real Time Enterprise
- Real Time Business Intelligence
- Information Integration
- Integration Brokers
- Enterprise Application Integration
- Operational BI
- Executive Dashboards
- Operational BI
- SCM, ERP, CRM
- Balanced Scorecard
- Business Activity Monitoring
- Zero Latency Enterprise
- Real Time Enterprise
- Real Time Business Intelligence
- Information Integration
- Integration Brokers
- Enterprise Application Integration
- Operational BI
- Executive Dashboards
Business Performance Management (BPM) is

- A business strategy
  - that aligns operations with corporate strategy
- for providing real-time access
  - increase velocity of business processes
- to critical business performance indicators
  - methodology for running the enterprise
- to improve business operations
  - for competitiveness, efficiency, lower costs, control
BPM – the Sum of Many Parts

Business Performance Management

Business Intelligence
Business Process Analysis & Modeling
IT Operations (NSM)

Enterprise Information Integration
Enterprise Application Integration

Data Warehouse
Production Applications
IT infrastructure
How the Data Warehouse Supports BPM

- **Context**
  - Compared to what? What happened last year? Last week?
    Forecast versus actuals?
  - Is this abnormal? Common?
  - When do we escalate?

- **Operational Data Store**
  - Often used in real time analytics & BPM

- **Analysis [filtering]**
  - Summaries & aggregates
  - Mining for patterns, predictions

- **Complex data integration**
  - Data cleansing & transformations
  - Shared data from multiple applications
Business Intelligence Best Practice

How do I this? How have others done it?
What’s the best way to achieve THIS?
Business Intelligence Best Practice: an IBM perspective

- Every Analyst and vendor has a list of Best Practices
- This material will try to be technology independent
- Business and Project Best Practices
- Design and technology Best Practices

Based on experience rather than theory
Customers the best practice is derived from:

- Vero Insurance New Zealand
- Standard Chartered Bank – Hong Kong and Singapore
- Bank of China (Hong Kong)
- BCB Malaysia
- TelstraClear New Zealand
- Korea Telecom
- Industrial and Commercial Bank of China
- Sprint PCS USA
- Axa SunLife: UK
- Telecom New Zealand
- Norwich Union UK
- Bank of Scotland UK
- National Australia Bank
- Lucas Aftermarket Operations
- British Steel (Chordiant)
Business Intelligence Best Practice: an IBM perspective

- Every Analyst and vendor has a list of Best Practices
- This material will try to be technology independent

- Business and project Best Practices
- Design and technology Best Practices
Business and Project Best Practice

- **Outsourcing and BI**
  - Control and stability vs unpredictability and change

- **Data Quality**
  - The value of “bad” data

- **Single view of customer**
  - Do you hold up loading the DW until can match/merge customers?

- **Profitability**
  - Where does the COST information come from

- **Incremental build plans**
  - Incremental means first deliverables are incomplete…

- **Industry Data Model**
  - Processes and analyses differentiate organisations, not the data structures
Business and Project Best Practice: Outsourcing and BI – some successes some limitations

- Keep knowledge in-house: you cannot discover and exploit new relationships and trends if you don’t understand the data
  - Data in the DW depends on ETL: you have to understand how ETL changes data to understand and exploit the DW
  - Data in OLAP systems, BI Applications and reports depends on ETL, Data Model and post-load processing… if you don’t understand these rules and changes you cannot exploit the information

- Change!
  - DW inherits the impact of change from every source system
  - Outsourcing is about minimising change
  - ETL and DW requirements WILL lead to changes in the source systems: how will your Outsourcer respond?
Business and Project Best Practice: Outsourcing and BI – which roles can be outsourced?

- Where is the boundary between mechanisms (outsource) and data/business knowledge (keep in house)?

<table>
<thead>
<tr>
<th>Role</th>
<th>Customer or Outsourcer?</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>DW Technical Architect</td>
<td>Customer</td>
<td>Strategic architecture requirements. Balances outsourcers drive to re-use technology even if not appropriate for BI/DW</td>
</tr>
<tr>
<td>Change Control Manager</td>
<td>Customer</td>
<td>Only the Customer understands true benefit of a proposed change. Impact definition is wider than outsourcer domain.</td>
</tr>
<tr>
<td>DW Server H/W and S/W support</td>
<td>Outsourcer</td>
<td>Implements to support DW Technical Architect design. Ensures DW uses common infrastructure to minimise costs.</td>
</tr>
<tr>
<td>Source Data Analyst</td>
<td>Outsourcer</td>
<td>Runs the source system(s): understands life cycle of data and events, how to extract. Assesses cost of proposed changes</td>
</tr>
<tr>
<td>Reporting and Analytics Analyst</td>
<td>Customer</td>
<td>Requires great flexibility and knowledge of the DATA</td>
</tr>
<tr>
<td>ETL Designer</td>
<td>Customer</td>
<td>Defines how business rules are applied to change and consolidate data</td>
</tr>
</tbody>
</table>
Business and Project Best Practice: Data Quality

- Correcting data is outside of the DW and ETL scope
  - ETL and the DW report data exceptions
  - Data corrections at source system or DW will not reconcile with source
- There is significant value in analysing “bad” quality data
  - Fraud (eg Sprint)
  - Anti Money Laundering
  - A business process is BROKEN: assess the impact
- For transactions, events, changes…
  - Anything where “correction” is via a reversal and a new posting
  - These have happened: so they need to be visible to BI users
    - Payments or receipts
    - Calls
    - Network outages
  - Store and analyse the original as well as the “corrected” data
Business and Project Best Practice: Single View of Customer

- Often a driving need to build a DW
  - Consolidate customer data from many sources
  - Apply consistent definitions across business units and geographies
  - Enable commercial customer roll-ups, residential householding
- Need to ask these questions:
  - How will putting all this disparate customer data into a single database produce a single view of customer?
  - What process will match/merge de-duplicate incoming customer records?
  - How will the ETL process know to combine incoming customer records or to consolidate into an existing DW row?
  - Can ETL wait for last incoming customer data before STARTING customer merge/match?
- Customer Match/Merge should be outside the ETL process
  - Best Practice is to separately develop match/merge processing and cross-reference of source system to DW Customer_ID
  - Temporary outcome: once data is loaded to the DW, re-process looking to consolidate customer data
Business and Project Best Practice: Profitability

- Often a driving need to build a DW
  - Single place where all revenues and activities associated with a customer or supplier are available
  - Looking for changes in profitability over time
  - Looking for behaviours that predict a change in profitability
- Revenue & invoice data usually in first DW projects
- Profit = Revenue (by customer, product, etc) – Costs (by customer, product, etc)
- Where does the COST data come from?
  - Activity based costing system: a source for the DW
  - Allocations within the ETL process?
  - Post load processing that allocates costs to customers, products, suppliers, etc?
Business and Project Best Practice:
Incremental DW development: a business driven roadmap

- **Staged Implementation Roadmap**
  - Sequence driven by business needs and data availability
  - Infrastructure work diminishes over time
  - 90 day delivery phases
  - Break big phases in two

Source: Dr. Barry Devlin, *Data Warehouse from Architecture to Implementation*, Addison Wesley, 1997, pp 315-318
Business and Project Best Practice: Industry data models

- Save significant project time
  - Up to 50% of definition phase

- Aids incremental development
  - Unambiguous scope
  - Re-use of data concepts

- Common business language
  - Helps ETL, reconciliation

- Bridge from business to IT and from requirements to deliverables

Information Framework (IFW)
Interlinked Models and Methods –

FSDM, TSDM, ISDM (RSDM, Postal, Airline)
Industry Concepts and Classifications

BDW, TDW, IIW (RDW, PDW, ADW)
Data Structures and Summaries

Profitability, Churn, ALM, Risk,
Usage, CRM,...
Pre-defined OLAP Cubes and Reports
Banking Data Warehouse Model

The Banking Data Warehouse Model (BDWM) provides a logical Entity-Relationship model of an enterprise-wide central data warehouse.

Contains:
- Flexible System of Record
- Commonly-required Summaries
- Sample Analysis Schemas
- Feedback Area

Intended to be customized
- Approx 80% of Customer’s Requirements for a Central Warehouse

Low-level flexible model
- Generic
- “One-step from Physical”

Enterprise-wide Logical Model
- Entity-Relationship
- Consists of 50 Subject Areas, 725 Entities, 3987 Attributes
- Fully defined and documented
Insurance Information Warehouse Model: pre-built analyses and application interfaces

- Catalogue of 1900+ insurance business terms
- 50+ Analytical Subject Areas
- 230+ Business metrics (KPIs)
- 350+ entities and relationships
- 4200 attributes
- 50+ pre-defined fact tables

<table>
<thead>
<tr>
<th>Business Profitability</th>
<th>Underwriting Analysis</th>
<th>Claims Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intermediary Commission Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Persistency analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Predictive modeling (fraud detection, lapses, …)</td>
<td></td>
</tr>
<tr>
<td>Analytical CRM</td>
<td>Customer Segmentation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lifetime value Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Campaign Management and Performance analysis</td>
<td></td>
</tr>
<tr>
<td>Financial Reporting</td>
<td>MIS (Dasboard)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KPI reporting</td>
<td></td>
</tr>
<tr>
<td>Risk Management</td>
<td>Risk Analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Risk Monitoring</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stochastics Modelling</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ALM</td>
<td></td>
</tr>
</tbody>
</table>
Telecom DW Model: inbuilt business solution templates

Customer Relationship Mgt
- Campaign Analysis
- Cross Sell Analysis
- Customer Acquisition Analysis
- Customer Churn Analysis
- Customer Behavior Analysis
- Customer Complaints Analysis
- Customer Interaction Analysis
- Customer Delinquency Analysis
- Wallet Share Analysis

Profitability
- Sales Channel Analysis
- Customer Profitability
- Product Profitability
- Customer Lifetime Value Analysis

Financial and Operational Reporting
- Credit and Collections Analysis
- Financial Management Accounting
- Revenue Analysis
- Individual Credit Risk Profile
- Service Order Processing Analysis

Usage
- Network Usage Analysis
- Service Usage Analysis
Business Intelligence Best Practice: an IBM perspective

- Every Analyst and vendor has a list of Best Practices
- This material will try to be technology independent
- Business and project Best Practices
- Design and technology Best Practices
Business Intelligence Best Practice:
Design and technology best practice

- Use a Layered Design
- Event-driven ETL
- Query prioritisation – intelligent queuing and governance
- Archiving and purging
- SMP vs Clusters
- Technology Do’s and Don’ts
Layered Data Architecture
Multiple levels of granularity provides various users with appropriate level of detail

Data Architecture designed specifically for mixed workloads, query performance, rapid incorporation of new data sources, and deployment of new applications.

EDW is an architecture, Not a database

- Users access a level
- Multiple levels of granularity
- Data summarization appropriate to user need
- As data ages, it rolls up through the layers
  - More tables. Less data per table
The layered architecture enables concurrent loading, query, archive and maintenance without compromising query performance.
Event-driven ETL: no schedule, data processed as it is available

- Example of Sprint PCS: no schedule, no outages, no reorgs
- Concurrent workload:
  - 70 queries per second
  - 300 million CDRs inserted per day – via 4 INSERTERs per partition
  - 300 million CDR’s are archived each day – via 2 DELETERs per partition
  - Use Static SQL with host variables: buffered INSERTS
- Continuous background loading – no schedule
  - Daemon programs look for input data, initiate Child Instances
  - Same mechanism for delete/archive
  - Rolling background on-line backups
- Summary tables updated – never/rarely refeshed
- Only clean data goes from Raw to SoR to Summary
  - Incomplete data: columns go as far as they can
  - Users interested in incomplete/incorrect data
Intelligent Queuing and Governance

- Mixed workload requires workload balancing
- Evaluation and workload management prior to resource consumption
  - Define a data path for executives and emergency work
  - Define a data path for “as long as it gets done” work
  - Define rules or priority for the rest (80%+)
- Evaluation and adjustment during resource consumption
  - Identify priority and “just get it done” work, allotting appropriate resources
  - Adjust the rest up and down based on priority, current workload and time period
Data Archiving and Retrieval

- More important where regulatory & compliance reporting needs more history than Decision Support and BI
- Set threshold for active vs aged data
- Use Hierarchical Storage Management (HSM) to leave stubs in aged tables, move data to tape
- Aged data (ie candidate for archive) still available to SQL – HSM retrieves on demand
- Generally over stressed, aged data will not usually influence or change aggregation and summaries
- Be fast enough to keep up
When to use an (UNIX) SMP, when to use a (UNIX) Cluster

- **Use a single SMP**
  - Ease of use to program, & manage
  - No unique skills required
  - Great reusability & resale value
  - Growth is containable

- **SMP should handle the workload for 2 years**
  - New CPU upgrade every 2 years solves capacity problems (Moore’s Law – IBM Power5)

- **Use cluster when one or more business requirements exceed large SMP capability**
  - I/O, concurrent users, database size, CPU burn rate
  - Serious need for High Availability fail over
  - Granular pay-as-you-grow business requirement

- **Only choose a cluster if SMP is not enough or is more expensive over 3-5 years**
  - Where scale out is necessary
  - Requires clustering skills & operations procedures
  - If you choose clustering, keep symmetrical
Do’s – Best Practices

- Build balanced symmetrical DB configurations
  - Keep it simple
- Identify & maintain correct ratio of disks to I/O to Server to CPU to Memory
  - Once the BCU ratios are defined, stick with it
- Always design for High Availability (HA) in cluster configurations
  - Be ready to add HA failover if not deployed initially
- Set AIX & DB2 configuration parameters right
  - DB2 configuration advisor will help a lot
Do’s – Best Practices

- Fit the technology to the business requirements
  - When technology & people clash, people win
- Get Database partitioning done right first
  - Select good partitioning keys
  - Ensure partitions map to disks for full parallelism & workload balancing
- Build a solid development & test environment
  - Ensure symmetry between queries, data, and operations
- Establish categories of node types
  - Data nodes, load-coordinator nodes, non-DB2 nodes
Don’ts: the Road to Troubles

- Don’t assume OLTP & BI workloads are similar
  - Caching & I/O throughput are different causing differing cost emphasis
  - Be careful consolidating SAN or disk subsystems for cost reasons alone
  - OLTP storage decisions are based on cost per GB.
    For BI, focus on cost per I/O to disk
Don’ts: the Road to Troubles

- Don’t use too few storage controllers vs. # of disks
  - BI is often constricted on I/O throughput

- Don’t use large disk capacities to save cost
  - 36GB disks ideal for many BI workloads

- Don’t assign > 1 DB2 partition to 1 disk
  - Keep the spindle serving 1 request, less head movement
  - Multiple partitions will have multiple I/Os = too much head movement for one disk spindle
Summary and close

Messages to take away
IBM’s leadership, research, products and support
Data Warehousing current and new waves:

- Consolidation
- Real-time (Right-time) Business Intelligence
- Extending the Data Warehouse via Enterprise Information Integration
- Business Performance Management
Business Intelligence Best Practice: an IBM perspective

- Every Analyst and vendor has a list of Best Practices
- This material will try to be technology independent
- Business and Project Best Practices
- Design and technology Best Practices

Based on experience rather than theory
Introduction: Roy Elgar: IBM Software Group

Data Warehouse Architect
AsiaPacific Advanced Technical Sales

relgar@nz1.ibm.com
+64 4 576 8152 office
+64 21 488 143 mobile