KIMBALL GROUP
Consulting | Kimball University

Rethinking EDW In The Era Of Expansive Information Management

Ralph Kimball
Informatica
March 2011

Ralph Kimball
Informatica
March 2011

Trends Driving the Expansion Of The EDW in 2011

- The big business trends → big data → big expectations
  - Management appreciates role of information assets: drive revenue growth, conversion rate, cost reduction, asset efficiency (doing more with less)
  - Competitors are monetizing opportunities, management is noticing!
  - Social media analysis, especially sentiment analysis becoming mainstream
- The EDW has become operational:
  - Real time data drives intervention, feature development, strategic planning
  - Irresistible push to zero latency data delivery
- Delivery modes have exploded → BI everywhere
  - Smart phones
  - Tablets
  - PCs at work, on the airplane, at home
  - Delivery to operational/control user interfaces, portals, alerts, tweets

KIMBALL GROUP

The EDW Must Change

- Importance and critical scrutiny of the EDW is growing as organizations monetize the decisions made with their data
- The size and types of the data being fed to the EDW is exploding – much new "big data" is not suited for relational processing
- EDW must adapt to different architectural roles: source or target, centralized or distributed, batch or real-time
- Many new unfamiliar technologies are required to exploit the data and stay competitive – need to shift the mindset to design for the unknown
- IT management is faced with an unprecedented array of choices and not much time to evaluate

KIMBALL GROUP

Increased Scope of the EDW

- No longer a library for simple numeric transactions
- The EDW has become the resource for all types of data assets including new types of data
  - Data assets are major component of the balance sheet, replacing traditional physical assets of the 20th century
  - Widespread recognition of the value of data even beyond traditional enterprise boundaries
  - Widespread understanding of how to monetize data – which must be timely, integrated and trustworthy
- EDW is not only the platform for business intelligence but is the platform for analytics

KIMBALL GROUP

What Are The New Distracting But Serious Constraints On The EDW?

- Compliance, privacy, security across new types of data and via new delivery modes, especially mobile
- Secure retention and archiving of structured data
- Adherence to new communications standards
- New generation of analysts and decision makers not tolerant of poor data quality

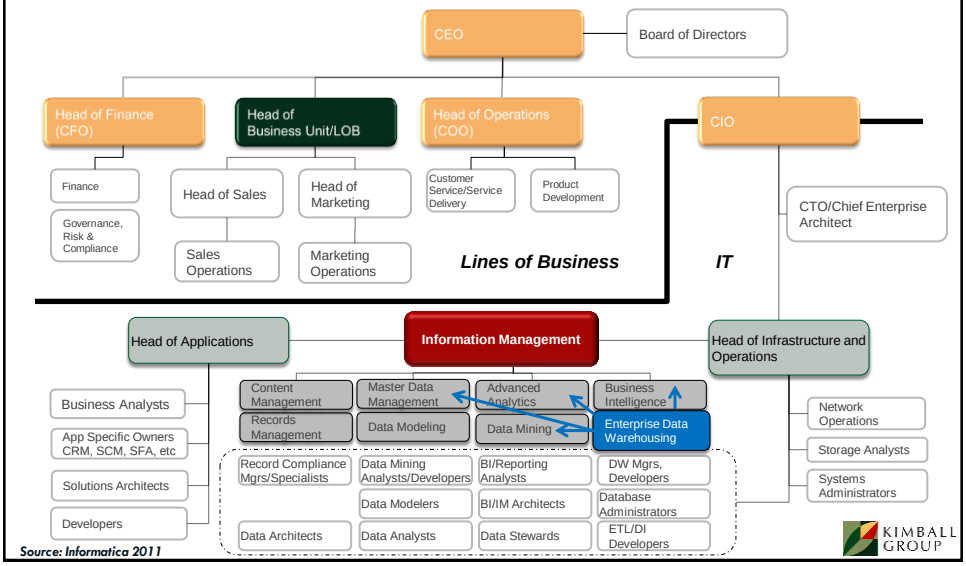
KIMBALL GROUP

EDW Must Drive the Evolving Analytic Capabilities in Information Management

- EDW MUST now support a broader set of analytic ecosystems:
  - Traditional data warehousing
  - Business intelligence
  - Master data management
  - Content management
  - Records management
  - Data mining
  - Data modeling
  - And, shared with business, advanced analytics
- Top level information managers are spending 50% or more time with the business

KIMBALL GROUP

EDW Team must Collaborate with Business and IT to enable a Portfolio of Projects



Source: Informatica 2011

What Analytic Demands Are Affecting The EDW?

- Social media processing requiring link analysis
- Increased sampling rates, Longer historical tails
- Need for unfocused full scans
  - Little willingness to statistically sample, thereby "losing the tail"
- Heavy need for non-relational analysis including iterative processing and complex branching logic
- Desire to process
  - Non scalar numeric data, even including waveforms
  - Non numeric "unstructured" data, including text, images and video
  - Extremely sparse, unpredictable "data bags"
- Complex event processing (CEP) affecting EDW
  - Retrospective batch analysis of EDW data with sophisticated data mining
  - Real time operational diagnosis and intervention, formerly bypassed EDW, now continuously streaming queries during load to EDW

KIMBALL GROUP

Data Bags: Extremely Sparse With Unpredictable Content

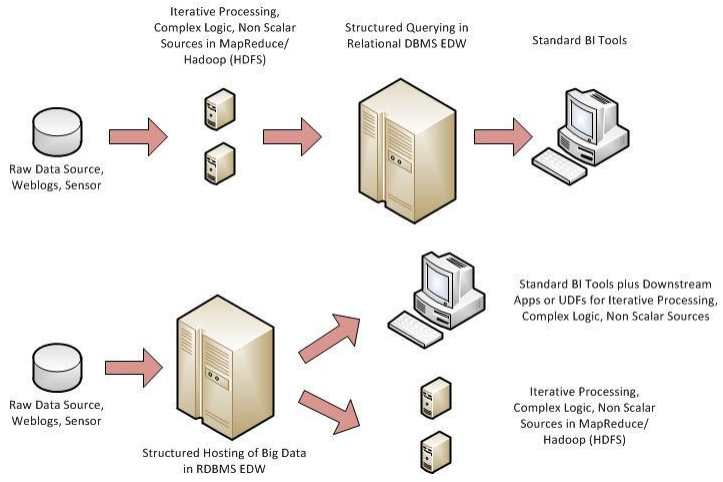
- E.g., financial disclosure information on loan application:
  - Payload is arbitrary list of name-value pairs
  - No limit to number of pairs or value data types
  - Data bags may contain data bags

Disclosure Fact
Application Date (FK)
Applicant (FK)
Loan Type (FK)
Application ID (DD)
Loan Officer (FK)
Underwriter (FK)
Branch (FK)
Status (FK)
Disclosure Payload (DB)

Disclosure Payload List Data Bag
Primary Income: $72345
Other Taxable Income: $2345
Tax-Free Income: $3456
Long Term Gains: $2367
Garnished Wages: $789
Pending Judgement Potential: $555
Alimony: $666
Jointly Owned Real Estate Appraised Value: $123456
Percentage Ownership Real Estate: 50
Number Dependents: 4
Pre-existing Medical Disability: Back Injury
Number Weeks Lost to Disability: 6
Previous Bankruptcy Declaration Type: 11
Years Since Bankruptcy: 8
Spouse Financial Disclosure: <data bag>
. . . And 100 more ...

KIMBALL GROUP

Relational EDW Can Be Either a Target or a Source for Analysis



KIMBALL GROUP

What Are The Most Serious Technical Hurdles For The New EDW?

- Ability to handle non-relational data including text, images, vectors, matrices, maps, and "data bags" of key-value pairs
- Ability to embed complex processing in database queries
- Extreme granularity: sub-transactional and light touch data
- Extreme distribution and extreme integration
- Extreme size: trillion row fact tables, billion row dim tables
- Big joins: joining these large tables across physically separated machines, not using clustered storage
- Pushing to zero latency, data creation to query ready
- Cumulative query processing, possibly stop at threshold
- Bandwidth failure: machine to machine network, cloud to cloud

KIMBALL GROUP

Database Architecture Alternatives

RELATIVE STRENGTHS	MPP Columnar RDBMS	MapReduce/ Hadoop
Unstructured data	Low	High
Complex logic	Low	High
Iterative processing	Low	High
Schema flexibility	Medium	High
Loading speed	High	Low
Maximum data size	Petabytes	Exabytes
Update, append	High	Low
Indexed lookup	High	Low
ACID Transactions	High	Low
Vendor SLAs (system)	High	Medium
Perceived Cost	Expensive	"Cheap" (open source)

KIMBALL GROUP

Software Architecture Alternatives for Big Data Analytics

- Embedded user defined functions (UDFs) in RDBMS
  - Subject to RDBMS API restrictions
  - Payloads as RDBMS "blobs", perhaps some SQL extensions
  - Processing on answer set, e.g., in BI layer
- MapReduce/Hadoop applications
  - Wider range of open source APIs
  - Payloads as fundamental file objects, full retrieval semantics
  - Map → Reduce processing at app top level
- Hybrid data movement, MapReduce/Hadoop → RDBMS
  - MapReduce/Hadoop hands off structured text/numbers to RDBMS after performing analysis and assembly of raw data
    - Finished and filtered analysis before handoff, or
    - Basic ETL processing, e.g.,
      - Sessionizing light touch data ocean
      - Massive sort before RDBMS data load

KIMBALL GROUP

Business/IT Organization Challenges

- CIO must become business relevant or die
- Unusual number of technical and architecture choices facing IT management
- Judging claimed feature convergence from vendors
  - objection removers, or true convergence?
- Analytics and data mining looking for a home
- Analyst antipathy to IT label
  - Shadow IT organizations sprouting within the business
- Analytics are naturally folded within the business organizations – need to be IT beachheads

KIMBALL GROUP

Business/IT Organization Responses

- IM (Information Management) subsumes DW, BI, MDM, records management, call it EDW
- Organizational model varies, but old school definitions dissolving:
  - DW in traditional IT
  - DW under Analytics in IT for broader IM
  - DW under Analytics in Business Unit/Technology Org.
- Develop analytics community identity and communication across business functions

KIMBALL GROUP

Interesting New Organization Developments Within IT

- Sandbox analytic environments with timeouts
  - Eclectic, non-standard tools
  - Conformed dimensions, shared data views
- Roll over sandboxes to production
  - Quick and dirty prototypes → standard implementations
  - IT teams up with analysts for roll over
- Agile development teams
  - Incremental development
  - Frequent releases
  - Small teams, possible end to end developer responsibility/involvement
- Cross functional analytics community
  - Newsletter, analytics portal, joint meetings, T-shirts & mugs

KIMBALL GROUP

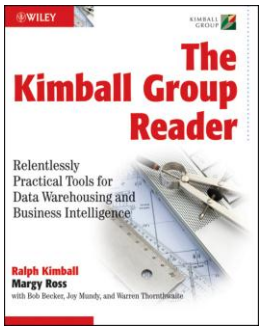
Immediate Steps

- Reassess analytic requirements
  from business needs perspective
- Expose alternatives, inventory skills sets
  - Database platforms, especially RDBMS and HDFS
  - System architecture
    - Where is data loaded, where does it go for processing
  - Software architecture
    - RDBMS or MapReduce/Hadoop or hybrid
- Develop criteria for build versus buy
  - Era of do-it-yourself system integration is almost over
- Educate management
  - agile approach, organization changes ahead

KIMBALL GROUP

The Kimball Group Resource

- www.kimballgroup.com
- Best selling data warehouse books
  NEW BOOK! The Kimball Group Reader ➔
- In depth data warehouse classes
  taught by primary authors
  - Dimensional modeling (Ralph/Margy)
  - Data warehouse lifecycle (Margy/Warren)
  - ETL architecture (Ralph/Bob)
- Dimensional design reviews and consulting
  by Kimball Group principals
- Informatica White Papers on Integration & Data Quality
- New Informatica White Paper on Big Data Analytics. Stay Tuned!



KIMBALL GROUP