NEWLY EMERGING BEST PRACTICES FOR BIG DATA

Ralph Kimball
Informatica
October 2012

Ralph Kimball

Big Data is Being Monetized

- Big data is the second era of data warehousing
  - First era (1980-2000): slice and dice transactions
  - Second era (2000 +): use analytics to tease insight from the massive universe of surrounding sub-transactions and new data sources
- Executives see the short path from big data insights to revenue and profit
  - Big data often illuminates behavior and preferences
  - Precise micro-marketing and micro-support drives cost savings and loyalty
  - Analytic sandbox results taken directly to management

Recognize Shocks to the System!

- RDBMSs and SQL can't store or process big data
  - Unstructured free text
  - Hyper structured data types, images, name-value pairs
  - Current system limits
- Shifting away from slice and dice reporting to analytics
  - Complex branching logic and iterations
  - Integration of diverse data types, historical, real-time
  - Dominated by full data scans, not indexed lookups
- Analysts continuously escalating requirements for lower latency analysis of petabytes of data

3

Newly Emerging Best Practices

- Big data showing first signs of maturity: general best practices becoming accepted
- To be useful we avoid motherhood and down-in-the-weeds
- Four best practice categories:
  - Management
  - Architecture
  - Modeling
  - Governance

4

Management Best Practices

- Structure big data environments around analytics, not ad hoc querying or standard reporting
  - Need exceptional freedom to define UDFs and construct arbitrary logic, processes, and objects
  - Leverage emerging technologies to complement existing
- Do not attempt to build a legacy big data environment at this time
  - Plan for disruptive changes: new data types, new algorithms, new hardware, new networking technology, new services
  - Reduce impact of disruptive changes with
    - Platform as a Service (PaaS)
    - Metadata driven environments

5

KIMBALL GROUP

More Management Best Practices

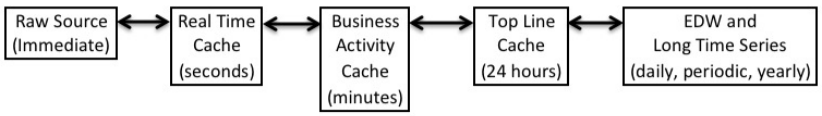
- Embrace sandbox silos and build a practice of productionizing sandbox results
  - Allow data scientists freedom to construct data experiments and build prototypes
  - After proof of concept systematically reprogram/reconfigure with an "IT turn over team"
- Put your toe in the water with a simple big data application: backup and archiving
  - Hadoop can be a low cost, flexible, format agnostic backup and archiving alternative

6

KIMBALL GROUP

Architecture Best Practices

- Plan for a logical "data highway" with multiple caches of increasing latency. Physically implement only those caches appropriate for your environment



  - Raw source: fraud detection, complex event processing
  - Real time: web page ad selection, personal promotions, game monitoring
  - Business activity: low latency KPI dashboards, trouble ticket tracking
  - Top line: quick review of last 24 hours, mid course corrections
  - EDW, long time: reporting, ad hoc querying, historical analysis, master data management
- Multiple paths from raw source with varying data completeness
- Important data flows in reverse directions

7

KIMBALL GROUP

More Architecture Best Practices

- Use big data analytics as a "fact extractor" to move data to the next cache
  - Unstructured data can be a rich source of structured dimensions and facts
  - For example, tweets can drive numerical, trendable sentiment measures including share of voice, audience engagement, conversation reach, active advocates, advocate influence, advocacy impact, resolution rate, resolution time, satisfaction score, topic trends, sentiment ratio, and idea impact
  - Investigate Informatica, Splunk and Kapow for extracting dimensions and facts from unstructured data

8

KIMBALL GROUP

More Architecture Best Practices

- Use big data integration to build comprehensive ecosystems that integrates conventional structured RDMS data, paper based documents, emails, and in-house business oriented social networking
- Use case (e.g., major brokerage house)
  - Millions of accounts, tens of millions of associated documents, and thousands of professionals both within the organization and also in the field as partners or customers
  - Set up a secure "social network" of all the trusted parties to communicate as business is being conducted
  - Capture all this information in Hadoop, dimensionalize it, use it in the course of business, and then back it up and archive it

9

KIMBALL GROUP

More Architecture Best Practices

- Plan for data quality to get better along the data highway
  - Latency trades off against quality
  - Data is fundamentally more complete further down the highway
- Apply filtering, cleansing, pruning, conforming, matching, joining, and diagnosing at the earliest touch point possible
  - Filtering, cleansing, pruning eliminates corrupted data
  - Conforming inserts highly administered standard descriptors
  - Diagnosing can insert confidence tags
- Implement backflows, especially from the EDW, to earlier caches on the data highway
  - Especially master data attributes for key dimensions
  - Reference data for lookups (e.g., useful keys, codes)

10

KIMBALL GROUP

More Architecture Best Practices

- Implement streaming data analytics in selected data flows
  - Serious analysis can be based on reaching thresholds during load
- Implement far limits on scalability to avoid "boundary crash"
  - Avoid a success disaster
- Do big data prototyping on a public cloud and then move to a private cloud
- Search for and expect 10x to 100x performance improvements over time, recognizing the paradigm shift for analysis at very high speeds
  - Performance improvements likely to involve new technology
  - Staying with Hadoop as the base is a good bet

11

KIMBALL GROUP

More Architectural Best Practices

- Exploit unique capabilities of in-database analytics
- Examples:
  - IBM's acquisition of Netezza and SPSS
  - Teradata and Greenplum's embedding of SAS
  - Oracle's Exadata R Enterprise
  - PostgreSQL's syntax for programming analytics and other arbitrary functions
  - Informatica's pushdown optimization to leverage in-database analytics as part of a data flow or ELT process

12

KIMBALL GROUP

Data Modeling Best Practices

- Think dimensionally: divide the world into dimensions and facts
  - Good application of big data analytics
  - Example tweet: "Wow! That is awesome!"
    - Extract customer (or citizen or patient), location, product (or service or contract or event), marketplace condition, provider, weather, cohort group (or demographic cluster), session, triggering prior event, final outcome
- Integrate separate data sources with conformed dimensions
  - Establish enterprise attributes with master data management, insert into early data steps
- Anchor all dimensions with durable surrogate keys

13

KIMBALL GROUP

More Data Modeling Best Practices

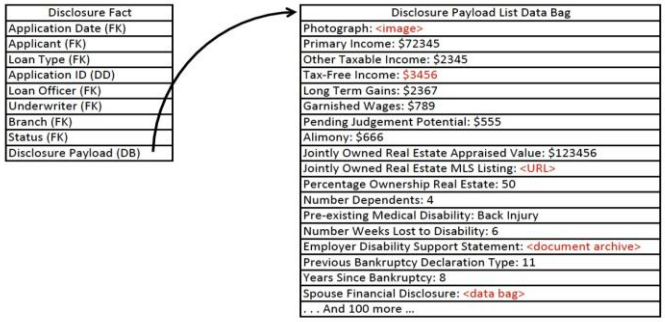
- Expect to integrate structured and unstructured data
  - Use data virtualization to insulate BI apps from underlying data changes, otherwise
  - BI tool must integrate results in last step
    (these are big architectural best practice, too)
- Track time variance with slowly changing dimensions (SCDs)
  - You have a duty to represent past history correctly
  - We REALLY know how to do this (SCD Types 1, 2, and 3)
  - See Kimball articles and books ☺
- Get used to not declaring data structures until analysis time
  - You may wish to change your mind or try alternatives
- Use data virtualization to allow rapid prototyping and schema alterations

14

KIMBALL GROUP

More Data Modeling Best Practices

- Build technology around name-value pair data sources:
  - Payload is arbitrary list of name-value pairs
  - No limit to number of pairs or value data types
  - Values can be arbitrary objects
  - Data bags may contain data bags



KIMBALL GROUP

Data Governance Best Practices

- There is no such thing as big data governance …
  - If you plan to ignore privacy, security, compliance, data quality, metadata management, master data management, and the business glossary, consider the business consequences
- Dimensionalize the data before applying governance
  - Yes, you may start governance before you understand content
  - BUT your best leverage is to dimensionalize early: find the customers, citizens, patients, locations, employees, products, services, …
- If analyzing data sets including identifying information about individuals or organizations, privacy is the most important governance perspective
- Don't put off data governance completely in your rush to use big data

16

KIMBALL GROUP

Summary

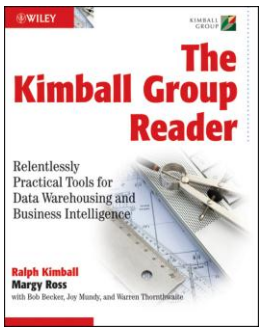
- We have a rich initial set of best practices
  - Management, architecture, modeling, governance
- Ignore these hard won lessons at your peril
- We have already passed by
  - Roll your own system integration
  - Program your MapReduce applications in Java
- Choose a flexible, changeable implementation platform: we are in the middle of dynamic changes of data types, analytic approaches, software
- Remember that this is still part of the EDW!

17

KIMBALL GROUP

www.kimballgroup.com Resource

- 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



- Kimball/Informatica White Paper expanding this webinar:
  http://needLink
- Foundation Kimball/Informatica White Paper on
  Big Data Impact on EDW:
  http://vip.informatica.com/?elqPURLPage=8808

18

KIMBALL GROUP

Big Data Best Practices

Analyze Outside the Box:
Turn Data Complexity Into Breakthrough Results

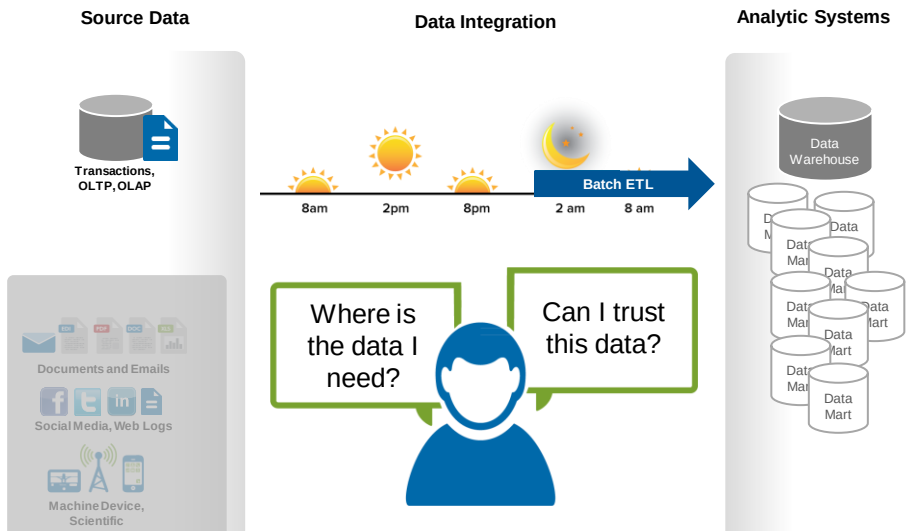
John Haddad
Director Product Marketing, Informatica

INFORMATICA

19

Big Data Challenges

Growing Volumes, More Variety, Lower Latency, Trust

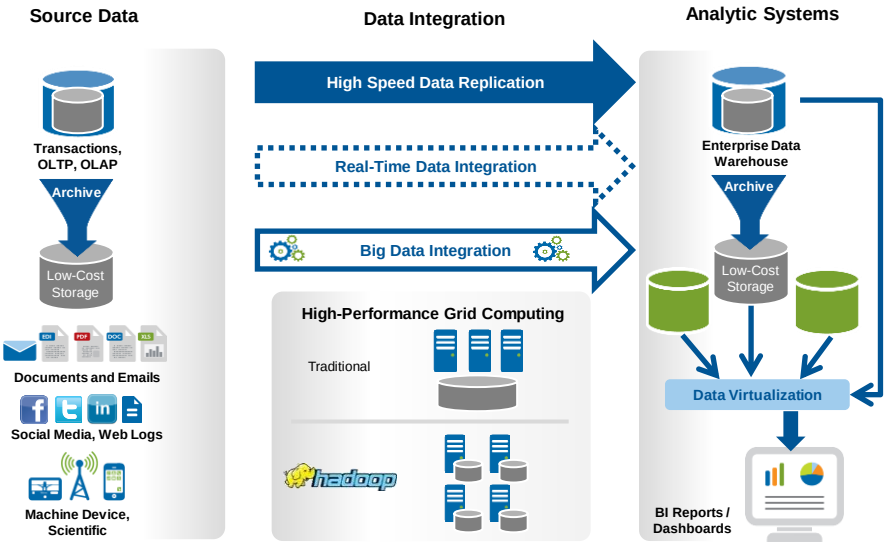


INFORMATICA

20

Big Data Integration

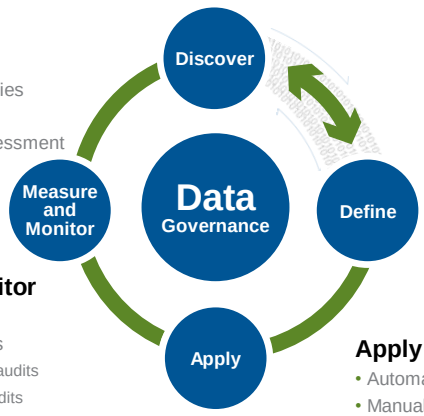
Lower Costs By Optimizing Storage & Processing



Data Governance

Improve business processes, decisions and interactions

Discover
- Data discovery
- Data profiling
- Data inventories
- Process inventories
- CRUD analysis
- Capabilities assessment



Define
- Business glossary creation
- Data classifications
- Data relationships
- Reference data
- Business rules
- Data governance policies
- Other dependent policies

Measure and Monitor
- Proactive monitoring
- Operational dashboards
  - Reactive operational DQ audits
  - Dashboard monitoring/audits
- Data lineage analysis
- Program performance
- Business value/ROI

Apply
- Automated rules
- Manual rules
- End to end workflows
- Business/IT collaboration

Big Data Best Practices Checklist

Cost-Effectively Manage the Volume, Variety, Velocity and Complexity

1. **Identify** inactive or infrequently used data and performance bottlenecks

2. **Archive** inactive data and stage raw data or infrequently used data to lower cost storage

3. **Execute** pre-processing and ETL on lower cost, scalable & reliable computing platforms

4. **Smooth** out ETL processing with real-time data integration

5. **Offload** processing from source systems with high-speed replication

6. **Eliminate** copies of data and augment the data warehouse using data virtualization

7. **Commit** to data governance to improve business processes, decisions and interactions

Next Steps

Kimball Corner

Listen to past webinars, download white papers, subscribe to Kimball Group *Design Tips*.

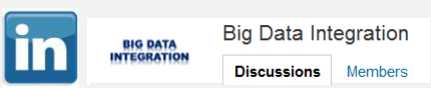
http://vip.informatica.com/?elqPURLPage=8007

Register for a no-cost Lean Data Warehouse Healthcheck

Includes a customized written assessment and consultation to identify dormant data and performance bottlenecks to optimize
http://vip.informatica.com/?elqPURLPage=9987

Join the conversation

LinkedIn Group:



For future webinars, demos, videos, whitepapers

www.informatica.com/bigdata