Extending Oracle’s Data Warehouse Reference Architecture for a Real Time and Big Data World

Peter Scott
Principal Consultant
Who Am I?

- One of the first members of Rittman Mead Consulting
- Specializes in Data Warehouse Design and Performance
  - Experience of several Realtime systems
- Writes and presents on many aspects of BI and Data Warehousing
Presentation Outline
Presentation Outline

• What is the Reference Architecture?

• Fitting in Realtime

• Going Large with Big Data
What is the Reference Architecture?

• Long established design methodology for building BI systems
  ▸ I have been using this technique since the 1990s

• Recognizes the needs of different data user communities and integrates that within a security framework
Components of a BI & Data Warehousing Architecture

- Storage
- Hardware & OS
- Database Server
- OLAP
- Query tools
- Metadata
- Security
- Ability to upgrade, replicate, clone, disaster recover
Major components

![Diagram of major components]

- Data Sources
- Information Sources
- Data Warehouse
- Information Management
  “Data Warehouse”
- ETL, Messaging and Metadata
- BI / PM Tools
- Information Access
- Security
Major components - data sources

Security

Data Sources
- Syndicated/External
- Unstructured Data
- Master Data
- Operational Systems
- COTS

BI / PM Tools

ETL, Messaging and Metadata

Information Management
“Data Warehouse”

Information Access
Data Warehouse Design - 3NF, Star Schemas or OLAP?

- Many DW designers take a “one or the other” approach to design
- Kimball-style star schemas are good for providing information access
- Inmon-style 3NF / CIF designs are good for storing and preserving data
- Ideal approach is to combine the two
  - 3NF “atomic storage layer” to preserve a process-neutral view of data
  - Star Schema “access and performance layer” to provide optimized access to data
Major components - information management
Major components - information access
What is the Reference Architecture?

• The reference architecture is extensible
  ▶ adding new data structures is simple
  ▶ model minimizes impact of change by allowing clear code and development paths
Fitting in Realtime

- Realtime is no different to any other data source
  - adding new data structures is simple
  - complexity is in ETL design to propagate change
    - Realtime changes may be ‘incomplete’ records so may need complex transformations in staging layer
    - Foundation to performance ETL will need thought to balance reporting needs and performance overhead
Major components - Realtime Integration
Major components - Realtime Integration
Realtime

• From Oracle’s February 2013 White Paper
Going Large with Big Data

• What is Big Data
  ‣ Unstructured???

• The “V”s of Big Data
  ‣ Velocity
  ‣ Variety
  ‣ Volume
  ‣ Value
  ‣ Veracity

• How is different to good ole’ data warehousing?
My take

• Big Data is about sentiment or “Aggregate meaning”
  ▸ Does it matter if we miss an event
    - “Yes” then this is transactional
    - “No” then this is Big Data

• Big Data is about storing volume and then statistically extracting data for later reporting
Going Large with Big Data

• Big Data is about sentiment or “Aggregate meaning”
  ‣ Does it matter if we miss an event
    - “Yes” then this is transactional
    - “No” then this is Big Data

• Big Data is about storing volume and then statistically extracting data for later reporting
  ‣ Map Reduce
  ‣ R
Going Large with Big Data

• Do we move “sentiment” into Data Warehouse?
  ▸ If “Yes” then output from Big Data storage is our source
    - Build ETL processes
    - Many Technologies
      - ODI Big Data Connectors
      - R
    - Code it yourself
Going Large with Big Data

• Keep “raw” big data out of the data warehouse

• But “reduced” data fits well

• Use reduced Big Data with foundation layer to enrich analysis

• Consider augmenting reporting tools with data exploration and visualization tools
Build a Sandbox to Explore Data
Build a Sandbox to Explore Data
Questions???
Presentation Takeaways

• The Data Warehouse Reference Architecture is a very flexible way of designing BI systems

• Adding Realtime data feed is hardly more complex than adding any other data source

• Adding Big Data is conceptually easy if you consider your Hadoop infrastructure as a source and map the reduced data into the Data warehouse staging area