From a Data Lake to a Data Hub

Allen Muller

May 8, 2018
A Little About Me…

Architect for:

• Data Hub – This presentation!
• Product Data Hub
• Trading Partner
• 18 Years at Eaton

And Other Random Stuff…

• Software Development Manager
• IoT and Machine Learning – in 1987!
• Nuclear Power Plant Simulation
• Survey Data Warehousing
The Eaton Journey from Relational to MarkLogic
Eaton At A Glance

• 96,000 Employees
• Global
• Power Management Focused DI
• Growth by Acquisition
We have a Complex BI Landscape

- Growth from $5B to $20B through acquisitions – many small
- 200+ Source Applications
- 19 Large Data Warehouses
- Until recently – 3 IT Organizations
- Many vertical solutions

Questions We Struggle With
- How will this hurricane affect our Supply Chain?
- What is our exposure to commodity tax changes?
- Who should we focus our marketing efforts on?

We Need to Centralize, Simplify, Hide Complexity, and Yes - Preserve Localized Data!
A New Architecture - First Pass

- Includes an “Integration Data Hub”
  - AKA ‘Conceptual ERP’ (CERP) – a view of Eaton
- Standard Technologies:
  - ETL, RDBMS
- 3rd Normal Form Canonical Data Model
- Started with One Subject Area
And the Results….?

- Variety and Volume of Sources Made CERP:
  - Difficult to design
  - Inflexible
  - Costly in terms of time, resources, money

- Resulting in:
  - Long delays for ‘Data In Use’
  - Missed opportunities
  - Inability to realize benefits

- The Idea is Good, the Tools are Wrong

Problems:
- Business logic in multiple places
- Tracing changes is VERY labor intensive

Problems:
- Brittle data model – RDBMS DDL
- Model changes ripple thru ETL and consumers

Problems:
- Localized data not accounted for – still need other ODSs or warehouses

Problems:
- Variety and Volume of Sources Made CERP:
  - Difficult to design
  - Inflexible
  - Costly in terms of time, resources, money

Problems:
- Brittle data model – RDBMS DDL
- Model changes ripple thru ETL and consumers

Problems:
- Localized data not accounted for – still need other ODSs or warehouses

Problems:
- Variety and Volume of Sources Made CERP:
  - Difficult to design
  - Inflexible
  - Costly in terms of time, resources, money

Problems:
- Brittle data model – RDBMS DDL
- Model changes ripple thru ETL and consumers

Problems:
- Localized data not accounted for – still need other ODSs or warehouses

Problems:
- Variety and Volume of Sources Made CERP:
  - Difficult to design
  - Inflexible
  - Costly in terms of time, resources, money

Problems:
- Brittle data model – RDBMS DDL
- Model changes ripple thru ETL and consumers

Problems:
- Localized data not accounted for – still need other ODSs or warehouses
Where ETL and RDBMS Struggled: Project-wise

**Inflexible Data Model Forces**
- **Choice:** Complex and extensible with full requirements – Time-consuming and costly, ‘cautious’ mindset
- **Choice:** Simple without all potential requirements – Done quickly but requiring expensive change later

**Changes Are Difficult**
- Synchronizing Releases: Harder as number of consumers grows

**Project Steps Must Be Addressed Sequentially**
A New Approach – Using the MarkLogic Database

Centralize
Transforms

Accelerate
Delivery

Add
Flexibility

Requirements

Plan

Deploy

Test

Develop

© 2018 Eaton. All rights reserved.
First: PoC to Compare MarkLogic to RDBMS

Constraints:

- Simple architecture
- Architect + Solution Designer
- Measure against previous project
- Two month effort
- Internal but leverage the MarkLogic team
MarkLogic versus RDBMS: Acceleration

- PoC efforts compared to Eaton Hydraulics and Aerospace for Invoices
- Factored in number of ERP variants
- Considered requirements gathering equal in both cases
- Added 50% to PoC time for ‘lack of formality’

Result: **4.2 to 1** acceleration using MarkLogic, based on work hours.
MarkLogic – Technical Advantages

- MarkLogic offers several technical advantages over traditional tools
  - Deployment Simplicity
  - Clustered Architecture
- MarkLogic capabilities extend the Data Hub’s flexibility
  - Semantics
  - Search
  - Many APIs – or build your own

Benefits:
- Flexible data model
- Envelopes: conformed and local data
- Data model versioning

Benefit:
- EL is much simpler than ETL

Benefit:
- Business transform logic all in one spot

Diagram:
- MarkLogic
  - Source
  - EL
  - Envelopes
  - Transform Logic
  - Source Documents
  - Reporting
  - Operational Systems
  - Advanced Analytics
  - Data Marts
Where MarkLogic Excelled – Project-wise

Flexible Data Model
- Removes ‘cautious’ approach
- Removes need for all requirements up front – low cost for changes

Data Model Versioning
- Decouple Data Hub activities from consumers
- Consumers can change when it makes sense for them

Flexible Data Model + Agile
- Vastly accelerated ‘Data In Use’

Consumer Change Projects
Just Delivered – Pilot – Inventory

- Put infrastructure in place – MarkLogic, SCC, etc.
- Build out MarkLogic teams
- Execute on a simple project
  - Note: MarkLogic no longer considered a ‘risk factor’ to management.
- Not expected to show acceleration here

1. Infrastructure
   - Infrastructure
   - Dev
   - QA
   - Prod
   - HA / DR
   - Process Docs
   - Validation

2. Platform
   - Software Platform
     - Install MarkLogic
     - Define Processes
     - Setup Support Tools
     - Process Docs
     - Validation

3. Application
   - Setup & Develop MarkLogic Application
     - Ingestion
     - Harmonization
   - Deliver OBIEE Reports
   - Process Docs
   - Validation
Next Up – Building a Program

• Three Initial Focus Areas
  • Finance Data
  • Manufacturing Sales and Orders
  • Supply Chain

• Challenges
  • Org Change Mgt
  • Agile
  • MarkLogic Skillsets
Summary

• For Eaton, A Data Hub Is Absolutely Necessary

• Traditional Tools Fall Short
  Technically
  Procedurally

• MarkLogic Offers Significant Benefits
  Flexibility of the Data Model
  Well-suited for Quicker Delivery Methods
  Canonical and Local Data Supported